AD-E100 093

LEVEL III

SPEECH QUALITY MEASUREMENT.

A052 272

By
T.P. Barnwell, III, A.M. Bush,
R.M. Mersereau and R.W. Schafer

School of Electrical Engineering
GEORGIA INSTITUTE OF TECHNOLOGY
Atlanta, Georgia

Final rept. 6 May 76–6 Jun 77,

E100 093

SBIE

FINAL REPORT E21-655-77-TB-1
Contract F30602-75-C-0118
10 June 1977

178 p.

Prepared For
DEFENSE COMMUNICATIONS AGENCY
DEFENSE COMMUNICATIONS ENGINEERING CENTER
1860 WIEHLE AVENUE
RESTON, VA 22090

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: E21-655-77-TB-1	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Speech Quality Measurement	5. TYPE OF REPORT & PERIOD COVERED: Final 6 May 1976 – 6 June 1977
6. PERFORMING ORG. REPORT NUMBER: E21-655-77-TB-1	
7. AUTHOR(s): T. P. Barnwell and A. M. Bush	8. CONTRACT OR GRANT NUMBER(s): F-30602-75-C-0118
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Georgia Institute of Technology, School of Electrical Engineering, Atlanta, GA 30332	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Post Doctoral Program, RADC/RBC (Mr. Jake Scherer), Griffiss AFB, NY 13441	12. REPORT DATE: June 10, 1977
	13. NUMBER OF PAGES: 162
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Defense Communication Engineering Center, 1860 Wiehle Avenue (Dr. W. R. Belfield), Reston, VA 22090	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Unlimited, Open Publication

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

Performed under the RADC Post Doctoral Program for DCA Defense Communication Engineering Center.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Digitization, Linear Predictive Coder, Voicing, Pitch, Quality Testing, PARM, Quart, DAM, Objective Quality Testing, Communicability, Subjective Testing, Acceptability Rating.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Speech quality measurement is considered from three points of view: subjective testing, objective testing, communicability testing. Speech quality is interpreted here in terms of user acceptability. It is assumed that good intelligibility is always present since otherwise a system is of no interest here.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Subjective testing is considered from the philosophical perspective of iso-preference, relative preference, and absolute-preference, with isometric and parametric test methodologies, with the results of PARM and QUART as a basis. It is felt that the best approach for future subjective testing will be a parametric approach using representative male and female talkers to cover the expected range of pitch. An automated and refined version of Voiers Diagnostic Acceptability Measure (DAM) test is an attractive option.

Objective testing is considered as a possible alternative to subjective testing. Reported here is a two part experimental study of the relationship between a number of objective measures and the subjective acceptability measures available from the PARM study. In the first part of the study, controlled distortions were applied to speech samples in order to measure the resolving power of the candidate objective measures on these types of distortions. In the second part, the candidate objective measures were applied to speech samples from the same systems on which PARM tests were run, and the statistical correlation between the objective and subjective measures were studied. Objective measures examined include spectral distance measures: Several LPC based spectral distances, LPC error power ratio, and cepstral distance; as well as pitch comparison measures, and noise power measures. Controlled distortions were formant bandwidth, frequency, pitch, low-pass bandwidth, and additive noise. Correlations with subjective test data range from ~ 0.2 to ~ 0.8.

In the communicability test, a somewhat different point of view is taken. The user is expected to perform on the data some cognitive task which is measurable. The rationale here is that the user will be better able to perform if the quality is high, than if his cognitive resource, assumed fixed, is saturated due to poorer quality transmission. The test format chosen for this study was a multiple digit recall test similar to that studied at Bell Labs by Naghtani. In this format sequences of random digits are first recorded by trained speakers, and then these utterances are played through various distorting systems. The resulting sequences are then played to subjects whose task is to "recall" the digits after a short ($\sim$ 1 second) wait. These tests prove to be rather unpleasant to take, and require larger numbers of subjects, but will differentiate among distorting systems.

FOREWORD

This effort was conducted by the School of Electrical Engineering under the sponsorship of the Rome Air Development Center Post-Doctoral Program for the Defense Communications Agency. Dr. W. R. Belfield of the Defense Communications Engineering Center was the task project engineer and provided overall technical direction and guidance.

The RADC Post-Doctoral Program is a cooperative venture between RADC and some sixty-five universities eligible to participate in the program. Syracuse University (Department of Electrical and Computer Engineering), Purdue University (School of Electrical Engineering), Georgia Institute of Technology (School of Electrical Engineering), and State University of New York at Buffalo (Department of Electrical Engineering) act as prime contractor schools with other schools participating via sub-contracts with the prime schools. The U.S. Air Force Academy (Department of Electrical Engineering), Air Force Institute of Technology (Department of Electrical Engineering), and the Naval Post Graduate School (Department of Electrical Engineering) also participate in the program.

The Post-Doctoral Program provides an opportunity for faculty at participating universities to spend up to one year full time on exploratory development and problem-solving efforts with the post-doctorals splitting their time between the customer location and their educational institutions. The program is totally customer-funded with current projects being undertaken for Rome Air Development Center (RADC), Space and Missile Systems Organization (SAMSO),

Aeronautical Systems Division (ASD), Electronic Systems Division (ESD), Air Force Avionics Laboratory (AFAL), Foreign Technology Division (FTD), Air Force Weapons Laboratory (AFWL), Armament Development and Test Center (ADTC), Air Force Communications Service (AFCS), Aerospace Defense Command (ADC), Hq USAF, Defense Communications Agency (DCA), Navy, Army, Aerospace Medical Division (AMD), and Federal Aviation Administration (FAA).

Further information about the RADC Post-Doctoral Program can be obtained from Jacob Scherer, RADC, tel. AV 587-2543, COMM (315) - 330-2543.

ABSTRACT

Speech quality measurement is considered from three points of view: subjective testing, objective testing, communicability testing. Speech quality is interpreted here in terms of user acceptability. It is assumed that good intelligibility is always present since otherwise a system is of no interest here.

Subjective testing is considered from the philosophical perspective of iso-preference, relative preference, and absolute-preference, with isometric and parametric test methodologies, with the results of PARM and QUART as a basis. It is felt that the best approach for future subjective testing will be parametric approach using representative male and female talkers to cover the expected range of pitch. An automated and refined version of Voiers Diagnostic Acceptability Measure (DAM) test is an attractive option.

Objective testing is considered as a possible alternative to subjective testing. Reported here is a two part experimental study of the relationship between a number of objective measures and the subjective acceptability measures available from the PARM study. In the first part of the study, controlled distortions were applied to speech samples in order to measure the resolving power of the candidate objective measures on these types of distortions. In the second part, the candidate objective measures were applied to speech samples from the same systems on which PARM tests were run, and the statistical correlation between the objective and subjective measures were studied. Objective measures examined include spectral distance measures: Several LPC based spectral

distances, LPC error power ratio, and cepstral distance; as well as pitch comparison measures, and noise power measures. Controlled distortions were formant bandwidth, frequency, pitch, low-pass bandwidth, and additive noise. Correlations with subjective test data range from ~ 0.2 to ~ 0.8.

In the communicability test, a somewhat different point of view is taken. The user is expected to perform on the data some cognitive task which is measurable. The rationale here is that the user will be better able to perform if the quality is high, than if his cognitive resource, assumed fixed, is saturated due to poorer quality transmission. The test format chosen for this study was a multiple digit recall test similar to that studied at Bell Labs by Naghtani. In this format sequences of random digits are first recorded by trained speakers, and then these utterances are played through various distorting systems. The resulting sequences are then played to subjects whose task is to "recall" the digits after a short ($\sim$ 1 second) wait. These tests prove to be rather unpleasant to take, and require larger numbers of subjects, but will differentiate among distorting systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Task History

The engineering effort reported on here was performed at Georgia Institute of Technology in the School of Electrical Engineering for the Defense Communications Agency through the Rome Air Development Center Post-Doctoral Program. The Post-Doctoral Program is under the direction of Mr. Jake Scherer. The monitoring officer at the Defense Communications Engineering Center was Dr. William R. Belfield, at the Defense Communications Engineering Center (DCEC).

This task, an investigation of subjective speech quality testing, objective speech quality testing, and communicability testing, was undertaken following the development at DCEC of a large data base associated with PARM and QUART (Paired Acceptability Rating Method and Quality Acceptance Rating Test). The existence of this data base has made possible the detailed analysis of subjective testing procedures, objective testing methods, and communicability testing, with good cross checking and validity referencing of results.

1.2 Speech Digitization Systems and Testing Requirements

Since it has for some years been clear that some form of end-to-end speech digitization would be initiated in the Defense Communication Systems, a number of speech digitization systems have been developed in various laboratories around the country. The job of selecting from these candidate systems the features to be included in a final system requires extensive evaluation and testing to be conducted. When a

"final" system is fielded, periodic field testing of all links for continued operational quality will be a significant requirement. This study attempts to further focus efficient means for developmental and operational quality testing.

1.3 Personnel, Procedures, and Facilities

This task has been carried out principally by Dr. T. P. Barnwell, with Dr. A. M. Bush, and with the active involvement of Dr. R. W. Schafer and Dr. R. M. Mersereau. Student Assistants have included Mr. Ashfaq Arastu, Mr. Bartow Willingham, and Mr. J. D. Marr here at Georgia Tech. This group also consulted on two occasions with Dr. W. D. Voiers of Dynastat, Inc., Austin, TX. The project was done for and with the active help of Dr. William R. Belfield of the Defense Communications Engineering Center.

Team leader was Dr. T. P. Barnwell. The project was initiated in May 1976 and completed in May 1977. Although six months effort was originally estimated, unavoidable delays in establishing the PARM data base at Georgia Tech delayed its progress. This report was prepared at Georgia Tech, tentatively approved in rough draft form at DCED, and subsequently reproduced at Georgia Tech.

This work was carried out in the School of Electrical Engineering Digital Signal Processing Facility. A block diagram is given as Figure 1.1. A more detailed description of the facility is given in Appendix C.

1.4 Technical Organization

The work reported here had as its ultimate goal the development of efficient objective methods and tests for predicting user acceptance

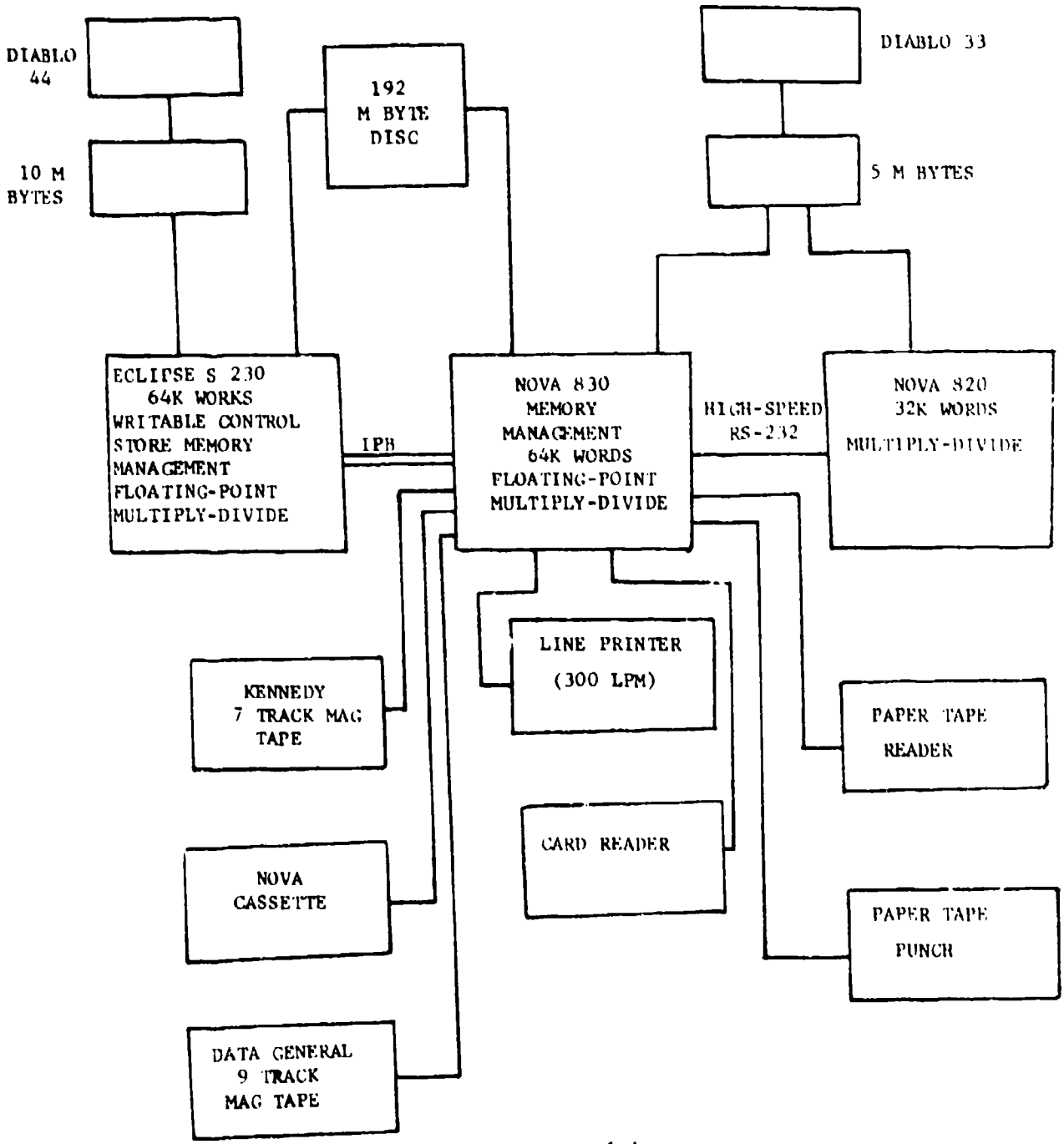


FIGURE 1.1

The Basic System for the Research Laboratory

of digital speech transmission systems. Three phases of the attack on this goal were established: (a) summary investigation of subjective testing methods; (b) development of a communicability test procedure; (c) development of objective testing procedures.

The outputs of the study are recommendations for future subjective test organization and implementation, specification of an objective testing procedure with cross-validation against PARM subjective testing results, specification of a communicability test philosophy and implementation of the test with results analyzed statistically. A secondary output is the PARM data base now organized for efficient searches.

Work progressed in all three phases in parallel, with some unexpected delays due to the time required to obtain and organize the data base from PARM (this is a large data base). A. M. Bush took principal responsibility for the subective testing portion, and T. P. Barnwell was principally responsible for the objective test and the communicability. R. W. Schafer and R. M. Mersereau also contributed to all three phases of the effort.

1.5 Organization of the Report

The detailed aspects of each of the three phases of the effort are presented in the report with the objective testing study in Chapter 2, the subjective testing study in Chapter 3, and the communicability test in Chapter 4. Each chapter is headed by an introduction giving the philosophy and rationale for that phase of the work and the technical perspective required for that phase.

II. OBJECTIVE MEASURES FOR SPEECH QUALITY

2.1 Introduction

In recent years, considerable effort has been devoted to the development and implementation of efficient algorithms for digitally encoding speech signals. These algorithms, which are utilized chiefly in digital communications systems and digital storage systems, cover a wide range of techniques, and result in systems which vary greatly in cost, complexity, data rate, and quality. Generally speaking, modern speech digitization systems can be divided into four categories: high rate systems which operate from ~ 100 KBPs to ~ 32 KBPs; intermediate rate systems which operate from ~ 32 KBPs to ~ 8 KBPs; low rate systems which operate from ~ 8 KBPs to ~ 1 KBPs; and very low rate systems which operate below ~ 1 KBPs. In the high rate systems, PCM [2.1] and adaptive PCM [2.2] are of the predominant techniques. In the intermediate rate systems, the techniques are more varied, including DM [2.3], ADM [2.4][2.5], DPCM [2.6], ADPCM [2.7], APC [2.8], and adaptive transform coding [2.9]. The low rate systems consist mostly of the vocoder techniques, including LPC [2.10-2.13], channel vocoders [2.14][2.15], phase vocoders [2.20][2.21], and several other techniques [2.22]. Very low rate systems usually involve feature extraction on a perceptual or linguistic level, and, thus far, very few systems of this type have been implemented. As a general rule, the higher data rate systems are less expensive to implement and less sensitive to bit errors, while the lower rate systems require more expensive terminals, and result in greater distortions in the presence of errors.

The problem of rating and comparing these systems from the standpoint of user acceptance is a difficult one, particularly since the candidate systems are usually highly intelligible. Hence, intelligibility tests, such as the DRT [2.23], may not suffice to resolve small differences in acceptability. Direct user preference tests such as the PARM [2.24] have been found useful for this purpose but are not highly cost effective. Moreover, they provide no diagnostic information which could be of value in remedying the deficiencies of systems being tested.

Objective measures which can be computed from sample speech materials offer a possible alternative to subjective acceptability measures. It should be noted, however, that the perception of speech is a highly complex process involving not only the entire grammar and the resulting syntactic structure of the language, but also such diverse factors as semantic context, the speaker's attitude and emotional state, and the characteristics of the human auditory system. Hence, the development of a generally applicable algorithm for the prediction of user reactions to any speech distortion must await the results of future research. However, the effects of certain classes of distortion are potentially predictable on the basis of present knowledge. In particular, substantial progress has been made in quantifying the importance of such acoustic features as pitch, intensity, spectral fidelity, and speech/noise ratio to the intelligibility, speaker recognizability as well as the overall acceptability of the received speech signal. Thus far, little success has accompanied efforts to predict the subjective consequences of other than relatively simple forms of signal degradation, but recent developments in digital signal processing techniques [2.25][2.26], suggest a number of efficient objective

measures which could be highly correlated with user acceptability.

In a recent study conducted by the Defense Department Consortium on speech quality, a large number of speech digitization systems were subjectively tested using the Paired Acceptability Rating Method (PARM) Test [2.24] developed at the Dynastat Corporation. The systems tested included a representative cross-section of the intermediate rate and low rate systems which had been implemented in hardware at the time of the study, and, consequently, offered a large user acceptability data base covering most classes of distortion present in modern speech digitization algorithms. The existence of the PARM data base offered a unique opportunity to measure the ability of objective measures to predict true subjective acceptability scores. Further, it allows the development of precise methodologies for the utilizations of objective measures in conjunction with subjective measures to possibly reduce the cost of speech system quality testing.

This chapter describes a two part experimental study of the relationship between a number of objective quality measures and the subjective acceptability measures available from the PARM study. In the first part of the study, controlled distortions were applied to speech samples in order to measure the resolving power of the candidate objective measures on these types of distortion. In the second part, the candidate objective measures were applied to speech samples from the same systems on which the PARM tests were run, and the statistical correlation between the measures, objective and subjective, were studied.

This entire chapter consists of five sections. In Section 2.2, the choice of objective measures is discussed. In Section 2.3, the "controlled distortion" experiment is presented. In Section 2.4, the

objective-subjective correlation experiment is described. Section 2.5 summarizes the results of this effort, and suggests directions for future research.

2.2 The Choice of Objective Measures

2.2.1 The Speech Perception Process

Human speech perception is a complex process in which distortions in the acoustic signal do not map simply onto perceived quality. In this section, several aspects of speech perception which relate to perceived speech quality will be discussed, and some general conclusions will be drawn.

First, it should be noted that the syntactic structure of a language has many components which impact speech perception. A sentence in a language may be viewed as a concatenation of phonemes which are hierarchically organized into syntactic and semantic units on a multitude of levels. Phonemes are grouped into syllables, syllables into words, and words into higher units (compounds, noun phrases, verb phrases, clauses, sentences, etc.) based on the phrase structure of the sentence [2.27]. Numerous modern linguists are trying to develop a comprehensive grammatical theory for the generation of the syntactical tree structures which represent the underlying sentence organization. The point here is that a great deal more information than the identity of the phonemes is being transmitted by the speech signal. Word boundaries, phrase boundaries, and many other syntactic elements have explicit correlates in the acoustics. It is these structural correlates which allow the listener to understand the sentence structure, hence, to use his great knowledge of the language to help him perceive the words themselves. Researches in speech synthesis by [2.28][2.29] have found

that the need to correctly produce the acoustic correlates of the syntax is at least equally important to correctly producing the acoustic correlates of the phonemes.

There is yet another level of information transmitted in the speech signal above the syntactic level. This level is semantic in nature, and incorporates the speaker's attitudes about the subject matter of the utterance. Linguistically, this information lies in the "intonation" and "emphasis" of the sentence, and this is also explicitly encoded in the acoustics.

When perceiving a sentence, a listener uses all these cues, phonemic, syntactic, and semantic, to help him understand the utterance. All these levels are highly redundant, and, in some cases, a great deal of acoustic distortion can occur without effecting the intelligibility or even the quality of the speech. However, in other cases, very slight distortions, such as those which effect the perception of syntactic structure, can cause complete loss of intelligibility. What is important in understanding the effect of a particular distortion is in understanding the way in which it interacts with the entire complex speech understanding process. At this point in time, even a simple complete enumeration of the information in a sentence is beyond the scope of current theory. This is why the problem of developing general objective quality measures is so difficult.

This is not to say, however, that there is not considerable knowledge about the acoustic correlates of the features of speech. It is well established that the phonemic information is primarily found in the acoustic filtering effect of the upper vocal tract, and hence, in the short time spectral envelope of the speech. Likewise, it is well

known that phase information, other than pitch, is not perceivable [2.22] Also, it has been well demonstrated that a great deal of information about consonantal identities are found in the formant behavior of the adjacent vocalics. But there are other phonemic acoustic correlates in English besides the spectral envelope. For example, voicing information in consonants is found in the durations of adjacent vowels and in the local pitch contour [2.30]

The major acoustic correlates of syntactic structure, intonation, and emphasis are pitch, vowel durations, and intensity. Of these correlates, pitch is by far the strongest [2.31] [2.32], followed by duration, and then intensity. There is also evidence that there are some effects in the spectral envelope which are involved in the perception of these "supersegmentals," though these are small.

When developing objective quality measures for intermediate rate and low rate digitization systems an important point is that, due to the nature of the systems themselves, only certain classes of distortions can occur. For example, phoneme durations, which are very important in perception of both phonemic and structural information, are not altered by coding. In vocoder systems, where the spectral envelope, pitch and excitation, and gain information are separated naturally as part of the digitization process, the mapping of the various parameters onto the perceptual domain is relatively easy to characterize. To detect distortion related to phonemic perception, spectral distance measures seem most important. Since the pitch contour plays such an important role in perception, some sort of excitation comparison should also be used. Since gain is relatively less important, it is expected that only gross gain errors should be detected.

In the case of waveform coders, the distortions are not so easily related to perception. Pitch information is not likely to be effected, but simple signal/noise ratios are not obviously good candidates for quality measures. A more likely candidate might be a measure based on the noise spectrum at the receiver.

2.2.2 Specific Objective Quality Measures

In this section, all of the objective quality measures tested in this study will be presented. All of the measures studied were not necessarily metrics. In order to qualify as a true metric, a distortion measure, D(X,Y), between two signals, X and Y, must meet the following conditions:

1. $D(X,Y) = 0$ iff $X=Y$
   $D(X,Y) > 0$ if $X \neq Y$

2. $D(X,Y) = D(Y,X)$

3. $D(X,Y) \leq D(X,Z) + D(Z,Y)$.

Some of the distortion measures in this study meet these requirements, while others do not.

2.2.2.1 Spectral Distance Measures

Spectral distance, in this context, refers to a distance measure between a sampled envelope of the source or unprocessed speech signal and a degraded form of the signal. Since there are many methods for approximating the "short time spectrum" of a signal, there are correspondingly many metrics which may be formed from a speech signal. A good measure should have two characteristics: it should consistently reflect perceptually significant distortions of different types; and, it should be highly correlated with subjective quality results.

A total of sixteen spectral distance measures and related

measures were studied in this project. Let $V(\theta)$, $-\pi \leq \theta \leq \pi$, be the short time power spectral envelope for a frame of the original sentence and let $V'(\theta)$ be the power spectral envelope for the corresponding frame of distorted sentence. In this discussion, it is assumed that the proper time synchronization has occured, and that $V(\theta)$ and $V'(\theta)$ are for the same frame of speech. Due to the fact the gain variations are not of interest here, the spectrums $V(\theta)$ and $V'(\theta)$ may be normalized to have the same arithmetic mean either in a linear or a log form. A geometric distance between the spectrums of the distorted and original spectrums may be taken in several ways, including direct spectral distance

$$D(\theta) = V(\theta) - V'(\theta) , \qquad 2.1$$

the difference in the log spectrums

$$D(\theta) = 10 \log_{10} V(\theta) - 10 \log_{10} V'(\theta) , \qquad 2.2$$

the source normalized distance measure,

$$D(\theta) = [V(\theta) - V'(\theta)]/V(\theta) \qquad 2.3$$

and the ratio of power spectrums

$$D(\theta) = V(\theta)/V'(\theta) . \qquad 2.4$$

Of these measures, 2.1 and 2.2 can form th basis for true metrics, while 2.3 and 2.4 cannot. A large class of distance measures can be defined as the weighted L_p norm "d_p" by

$$d_p(V,V',W) = \left[\frac{\int_{-\pi}^{+\pi} W(V,V',\theta) |D(\theta)|^p d\theta}{\int_{-\pi}^{+\pi} W(V,V',\theta) d\theta} \right]^{1/p} \qquad 2.5$$

where $W(V,V',\theta)$ is a weighting function which allows functional weighting based on either of the power spectral envelopes or on frequency. In this study, $W(V,V',\theta) = 1$, and 2.5 reduces to

$$d_p(V,V') = \left[\frac{1}{2\pi} \int_{-\pi}^{+\pi} |D(\theta)|^p d\theta \right]^{1/p} . \qquad 2.6$$

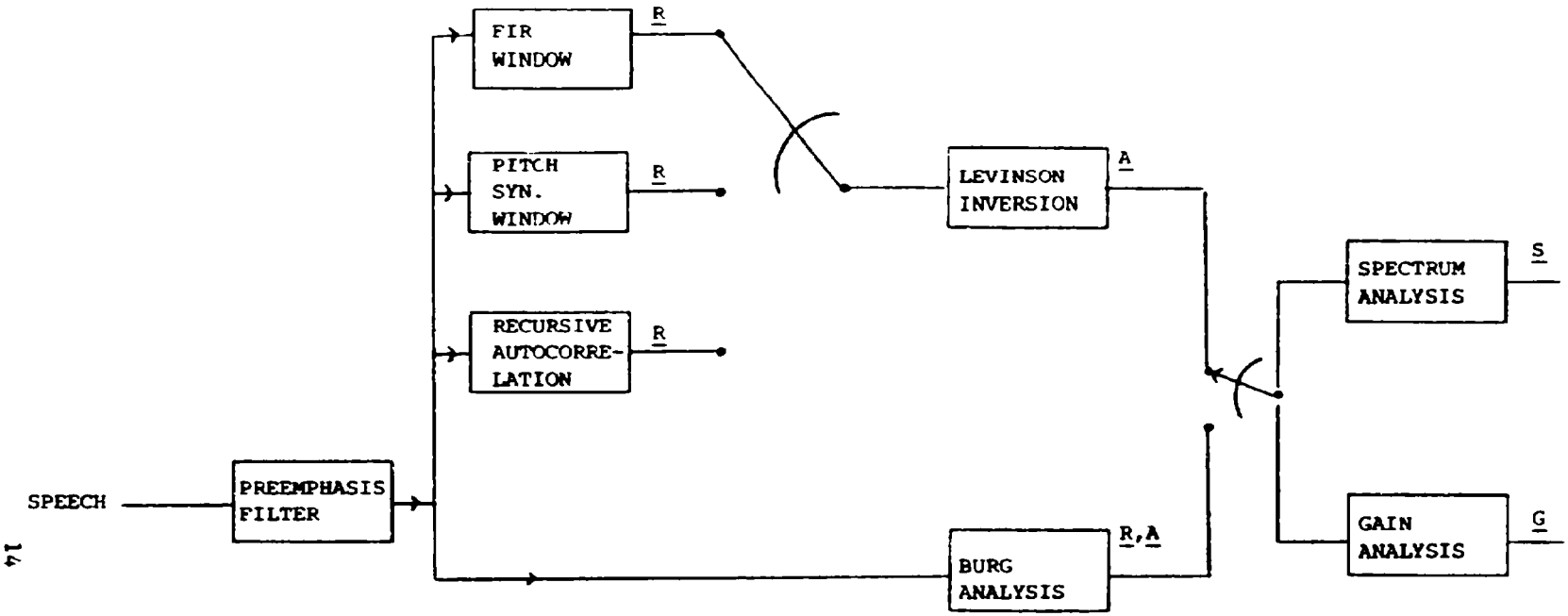
Clearly, the higher the value of "p," the greater the emphasis on large spectral distances. This measure may be digitally approximated by sampling $D(\theta)$, giving

$$d_p(V,V') \approx \left[\frac{1}{M} \sum_{m=1}^{M} |D(\frac{m\pi}{M})|^p \right]^{1/p} \qquad 2.7$$

2.2.2.1.1 The LPC Spectral Distance Measures

Since the output speech waveform is a convolution between a spectral envelope "filter" and excitation signal, then a deconvolution is necessary for spectral envelope comparisons. The LPC analysis is itself a parametric spectral estimation process, and may be used to extract an approximation of the spectral envelope. The block diagram for an LPC spectral analysis system is given in Figure 2.1. If the LPC parameters are $(a_1, \ldots, a_n)$, then the spectrum function $V(\theta)$, is given by

$$V(\theta) = \frac{G^2}{|A(e^{j\theta})|^2} \qquad -\pi \le \theta \le \pi \qquad 2.8$$



$$G = R_o - \sum_{i=1}^{N} a_i R_i$$

$$S_k = \left| \frac{1}{1 - \sum_{i=1}^{N} a_i e^{\frac{j\pi ik}{m}}} \right| \qquad M = \text{no. of samples}$$

LPC SPECTRUM ANALYSIS

FIGURE 2.1 SYSTEM TO DO LPC SPECTRUM ANALYSIS

where

$$A(z) = 1 - \sum_{i=1}^{N} a_i z^{-i} \quad . \tag{2.9}$$

This approximation can be used to calculate any of the measures suggested above.

There are a number of additional measures which can be calculated from A(z). These are not true spectral distance metrics or measures, but are related, and have the additional feature that they are easy to calculate. Several of these measures are simply geometric distances in the parameter domains, such as feedback coefficients, PARCOR coefficients, area functions, and pole locations. In each of these cases, we can define d_p as

$$d_p(\xi,\xi') = [\frac{1}{N} \sum_{m=1}^{N} |\xi_m - \xi'_m|^p]^{1/p} \tag{2.10}$$

where ξ_m is the m^{th} parameter (PARCOR coefficient, area function, etc.), and N is the number of parameters involved in the representation.

Another related approach is illustrated in Figure 2.2. The original speech signal is analyzed using an LPC analysis, and the inverse filtered waveform is formed by

$$e_i = s_i - \sum_{j=1}^{N} a_j s_{i-j} \tag{2.11}$$

where a_j is the j^{th} LPC coefficient and s_i is the i^{th} speech sample. This optimal filter is then used to inverse filter the distorted waveform, resulting in

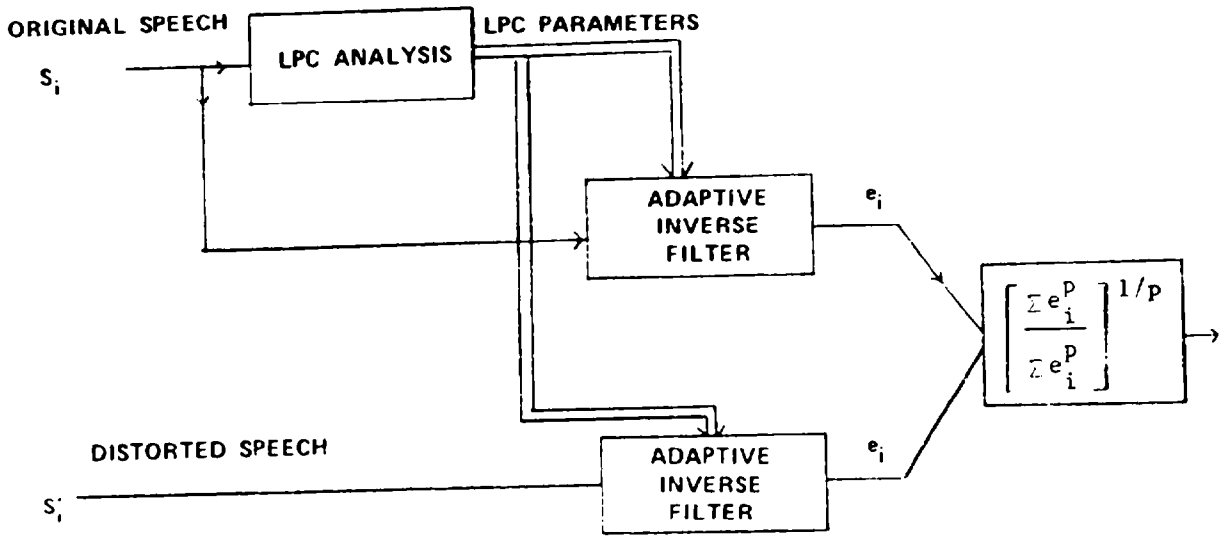


FIGURE 2.2 SYSTEM FOR COMPUTING THE "ERROR POWER RATIO" MEASURE.

$$e'_i = s'_i - \sum_{j=1}^{N} a_j s'_{i-j} \quad . \tag{2.12}$$

The measure which is used is then

$$d_p = \left[\frac{\sum_{i=1}^{L} e'^{p}_i}{\sum_{j=1}^{L} e^{p}_i} \right]^{1/p} \quad , \tag{2.13}$$

where L is the total number of samples in the utterance.

2.2.2.1.2 Cepstral Spectral Distance Measures

Another technique used often for deconvolving the spectral envelope from the excitation is cepstral analysis [2.33][2.34]. The analysis system for cepstral analysis is shown in Figure 2.3. By Parseval's Theorem, d_2 can be calculated from the cepstrum by

$$d_2 = \sum_{k=o}^{\infty} |c_k - c'_k| \tag{2.14}$$

where C_k and C'_k are the cepstral components for the original and the test signal respectively. For the same reason that cepstral deconvolution works well on speech, only a few coefficients need to be used ($\lesssim$ 40) to calculate d_2. Since the cepstral measure is computationally intensive (2 FFT's per frame) and since it has been shown that d_2 calculated from A(z) is very highly correlated with d_2 calculated from the cepstrum [2.35], then it does not appear that the cepstral measure is very attractive. However, the cepstral measure is attractive for excitation feature extraction (see 2.2.2.2.2); since the low order cepstral coefficients are a by-product of that analysis, and since CCD's

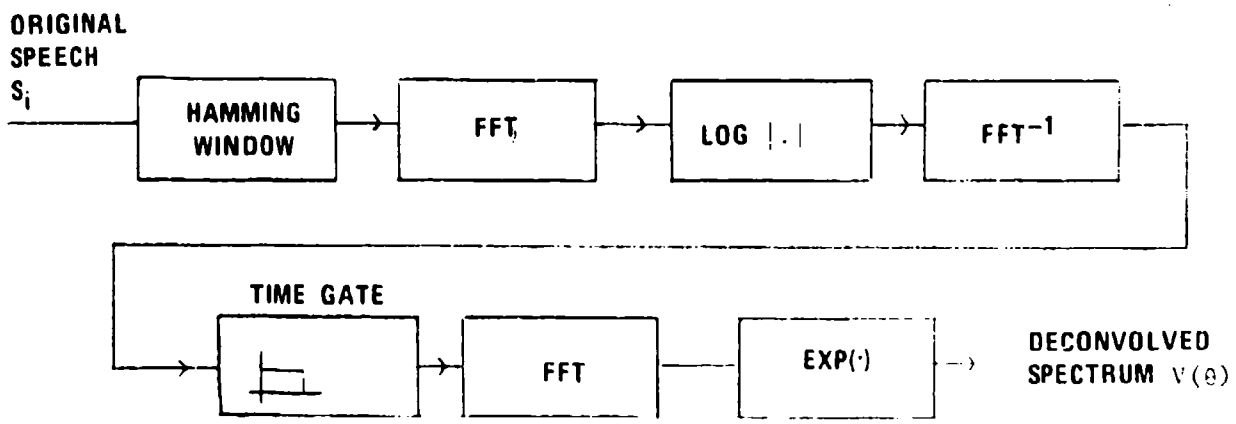


FIGURE 2.3 SYSTEM FOR PERFORMING CEPSTRAL DECONVOLUTION FOR SPECTRAL ENVELOPE ANALYSIS.

offer potential for cheap FFT's using the CHIRP-Z Transform, then cepstral measures are worthy of consideration.

2.2.2.2 Excitation Feature Extraction

Pitch is a very important acoustic correlate of many supersegmental features, and distortions in the pitch contour are easily perceivable and very detrimental to quality. Pitch estimation errors and voiced/unvoiced errors may occur in any pitch excited vocoder system. Hence, it is of interest to investigate objective measures for comparing excitation features for those systems where it is applicable.

The ideal solution to this problem would be to generate high quality pitch contours for the original utterances, and to compare these to the values used by the vocoder synthesis algorithm. However, since the excitation parameters are not explicitly available in vocoder systems, and since the excitation data is not available for the systems used in the PARM test, then this approach is unreasonable.

A second possibility is to apply a high quality pitch detector to both the original and the distorted speech, and to compare these results. A system which compares pitch excitation contours was developed at Georgia Tech under a previous effort [2.36] along with several high quality pitch detection programs. The statistics performed by the pitch comparison program (PCHECK) are enumerated in Table 2.1. This approach was studied experimentally using the Hard Limited Autocorrelation Pitch Detector [2.36] and the Multiband Pitch Detector [2.36].

A third possible approach involves developing a measure for excitation differences which does not depend on any pitch detection algorithm. The idea is to use a deconvolution technique which is aimed at retrieving the excitation representation rather than the spectral

STATISTICS
1. Total number of pitch errors
2. The average errors per sample in voiced regions
3. The number of gross errors (greater than a threshold)
4. The average gross errors
5. The number of subtle errors (less than a threshold)
6. The average subtle errors
7. The number of voicing errors
8. Sample standard deviations from the above averages

2.1 Statistics Calculated by "PCHECK" Pitch Comparison Program

envelope representation. The cepstrums of the two speech signals have many features which suggest that they might be good candidates for an excitation distance measure. First, they have a region in which the signal characteristics are almost entirely representative of the excitation function. Second, since this region is easily identifiable, no pitch decision or voiced/unvoiced decision is necessary. Third, the shape of the cepstrum in the excitation region contains some additional information about the excitation besides just pitch. Last, the computation of the cepstrum leads to a spectral envelope representation which might also be used as part of a spectral distance measure.

The way in which an excitation distance measure might be calculated is illustrated in Figure 2.4. After the cepstrum of the two signals is calculated, a smoothing filter is used to make the measure less severe. Next, a distance metric is calculated by

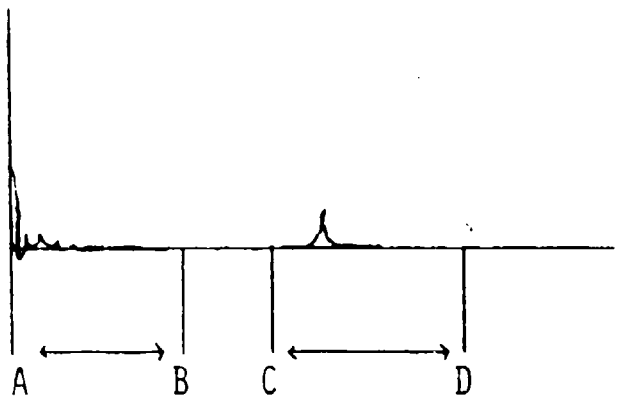
$$d_p = \left[\frac{\sum_{k=N1}^{N2} W(C,C',k)\,(C_k - C'_k)^p}{\sum_{k=N1}^{N2} W(C,C',k)} \right]^{1/p} \quad . \qquad 2.15$$

In this measure, C_k and C'_k are the cepstral coefficients for the original and distorted speech respectively, and $W(C,C',k)$ is a weighting function. In this study, the weighting functions which were studied were $W(C,C'_j,k) = 1$ (no weight) and $W(C,C',k) = C_k$, which weights samples near pitch peaks more than those in unvoiced regions.

2.2.2.3 Noise Power Measures

Traditionally, signal-to-noise ratio has been one of the predominant measures for determining the performance of waveform coding

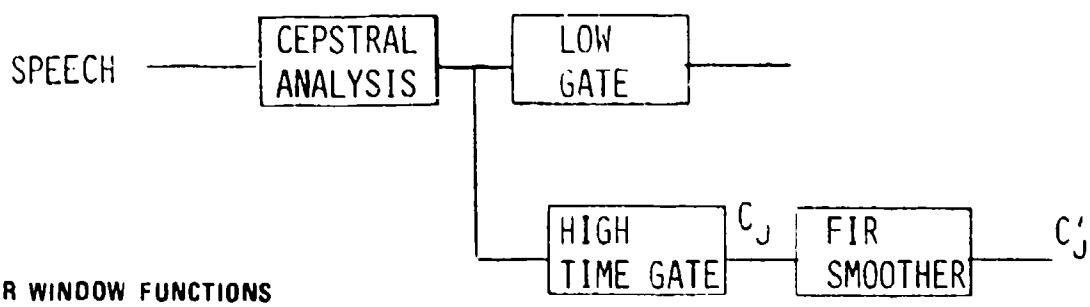
CEPSTRAL EXCITATION METRIC



EXCITATION INFORMATION

SPECTRAL
INFORMATION

$$D_i = \sum_{J=C}^{D} \left| C_J^M - C_J^S \right|$$



FIR WINDOW FUNCTIONS

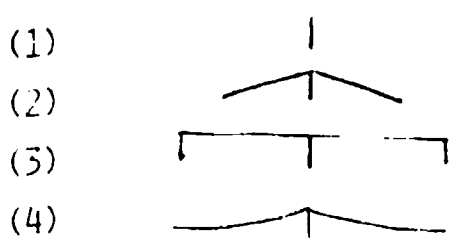


FIGURE 2.4 CALCULATION OF THE CEPSTRAL PITCH METRIC

systems. This measure is attractive since it is so easily calculated and since values for this measure are known for most appropriate systems. It is unattractive since it is difficult to evaluate in light of what is known about speech perception.

A far more interesting approach might be to develop a measure based on the coloration of the noise as well as its power. In short, if noise is defined as

$$n_i = s_i - s'_i \quad , \tag{2.16}$$

where s_i and s'_i are samples of the original and distorted speech respectively, then the noise spectral envelope $N(\theta)$ could be found using LPC or cepstral techniques as before. A measure could be defined such that

$$n_p = \frac{\int_{-\pi}^{+\pi} W(\theta) N^p(\theta) d\theta}{\int_{-\pi}^{+\pi} W(\theta) d\theta} \tag{2.17}$$

and

$$d_p = \left[\frac{\sum_{i=1}^{n} s_n^p}{n_p} \right]^{1/p} \tag{2.18}$$

This would be attractive since it would allow some measure of the spectral characteristics of the noise, which is very likely to have perceptual impact. If $W(\theta)=1$, then, by Parseval's Theorem, this measure becomes the signal-to-noise ratio for $p=2$.

Though this represented a very interesting area for study, very

little was done on noise measurements in this study. This is because the data base associated with the DARM was not in a form to make the necessary computations reasonable.

2.3 Initial Qualitative Studies and Controlled Distortions

This section describes two phases of the experimental study. In the first phase, example sentences from various systems were digitized from analog magnetic tape, and various forms of gain measures and spectral measures were applied and studied. In the second phase, the measures presented in the previous section (2.2) were applied to sentences which contained controlled distortions to test these measures for consistency in measuring these distortions, to check the measurement of combined distortions, and, by using the histograms of time behavior of the various measures, to determine a potential resolving power for each measure.

2.3.1 Qualitative Studies

In the initial study, a total of 20 sentences from two speakers and five systems were digitized from analog tape (digital tape representations were not available at that time), and stored on disk. (See Table 2.2.) A subgroup of those sentences was then analyzed for energy contours and for spectral representations and cepstral spectral analysis.

The energy was measured by applying Kaiser windows [2.37] of various lengths as FIR filters to the squared waveforms. The window lengths were adjusted such that pitch periods were not obvious in the energy representations. These energy plots were then used to try to synchronize the sentences with one another.

Several results came out of this study. First, not unexpectedly, the energy plots for the waveform coders (CVSD 16 and CVSD 9.6) were

TEST UTTERANCES

HI ANCHOR	LL1*	LL2	CH1	CH2
CVSD (16 KBPS)	LL1*	LL2	CH1	CH2
CVSD (9.6 KBPS)	LL1*	LL2	CH1	CH2
LONGBRAKE (2.4 KBPS)	LL1*	LL2	CH1	CH2
HY2 (2.4 KBPS)	LL1*	LL2	CH1	CH2

*Part of Subtest Group

Table 2.2 Input Sentences Used in the Initial Qualitative Studies

very similar to that of the high anchor (original). Second, the energy plots for the vocoders (Longbrake 2.4 and Hy2 2.4) were very different from the high anchor and very different from each other. Attempts to synchronize the utterances using the gain waveforms result in different synchronizations than if the waveforms are synchronized visually. The point here is that since the local intensity of a speech waveform is not a highly perceivable quantity, and vocoders take advantage of this by doing relatively poor gain estimation, and points out that energy is probably not a good candidate for an objective quality measure.

Another point should be made here. The synchronization efforts here point up clearly that the use of analog magnetic tape for recording utterances is generally unacceptable. Effects which (we presume) are due to the stretching of the analog tapes prevented synchronization from being maintained for more than 1-2 seconds. Carefully synchronized digital playback and recording systems must be used as a basis for reasonable objective measures.

In the second part of this study, 10 pole LPC spectral analysis and 40 coefficient cepstral spectral analysis was performed on the five test sentences, and 3-D perspective plots were produced. These plots are shown in Figures 2.5-2.14. Several points were observed from these plots. First, the peaks in the LPC spectra were generally sharper than those of the cepstral spectra. Second, however, the cepstral spectra, on the whole, had much more local variations than the LPC spectra. Third, the spectral variations caused by the waveform coders were more noticable in the LPC case than in the cepstral case. On the whole, no clear advantage for either of the two analyses could be found from these plots.

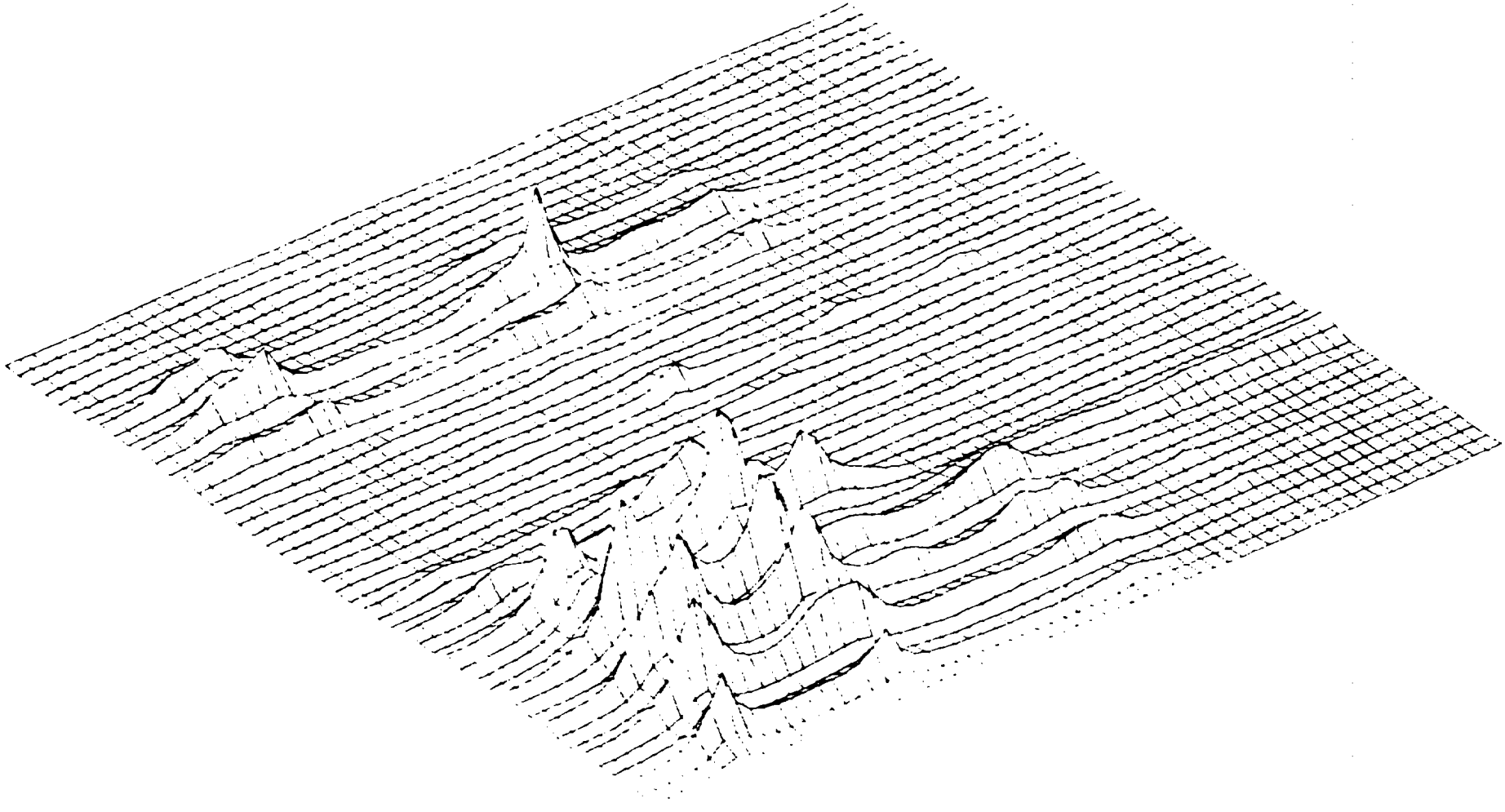


FIGURE 2.5 LPC SPECTROGRAM OF HIANCHOR (LL1)

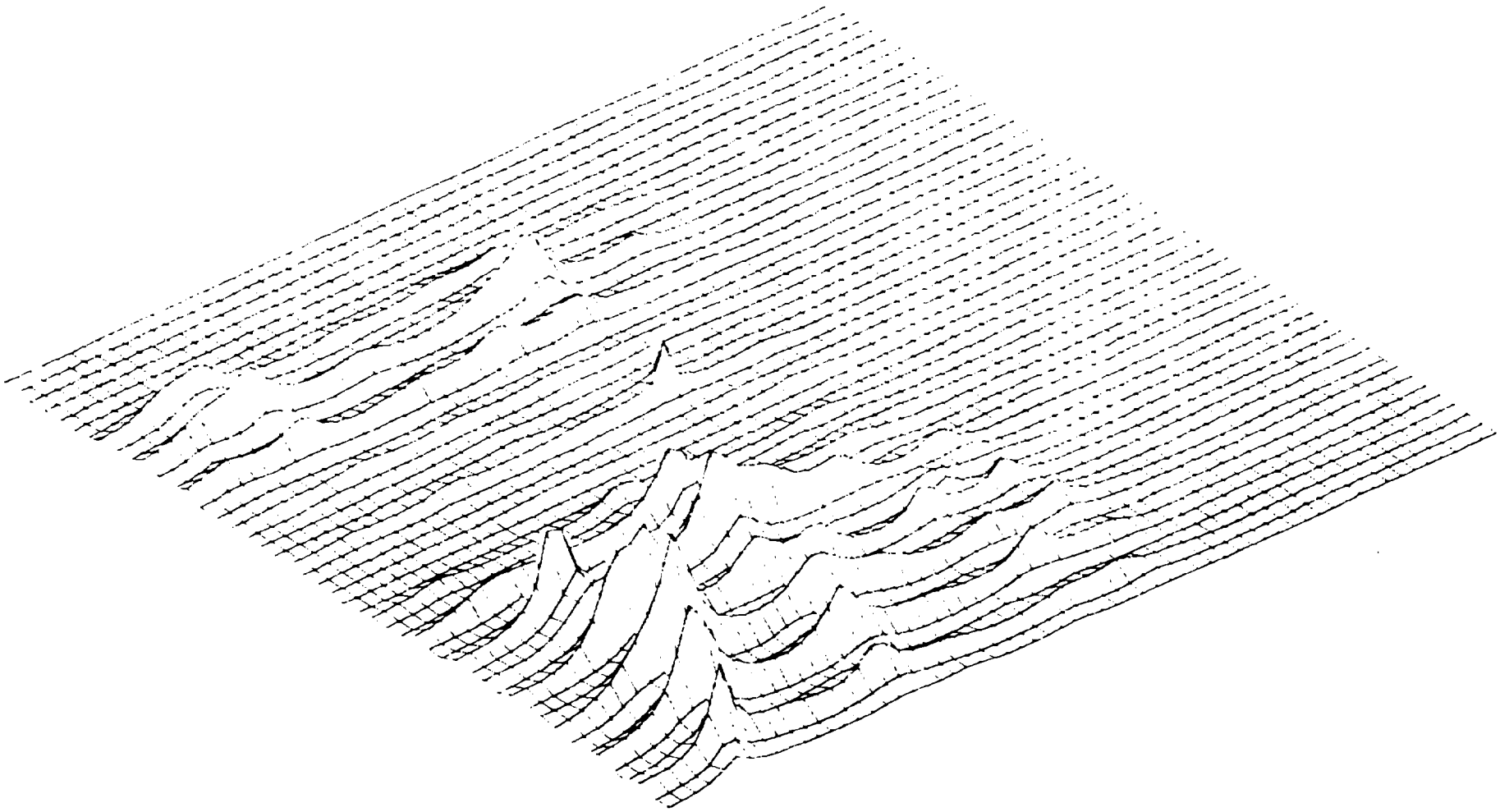


FIGURE 2.6 LPC SPECTROGRAM OF CVSD AT 9.6 KBPS (LL1)

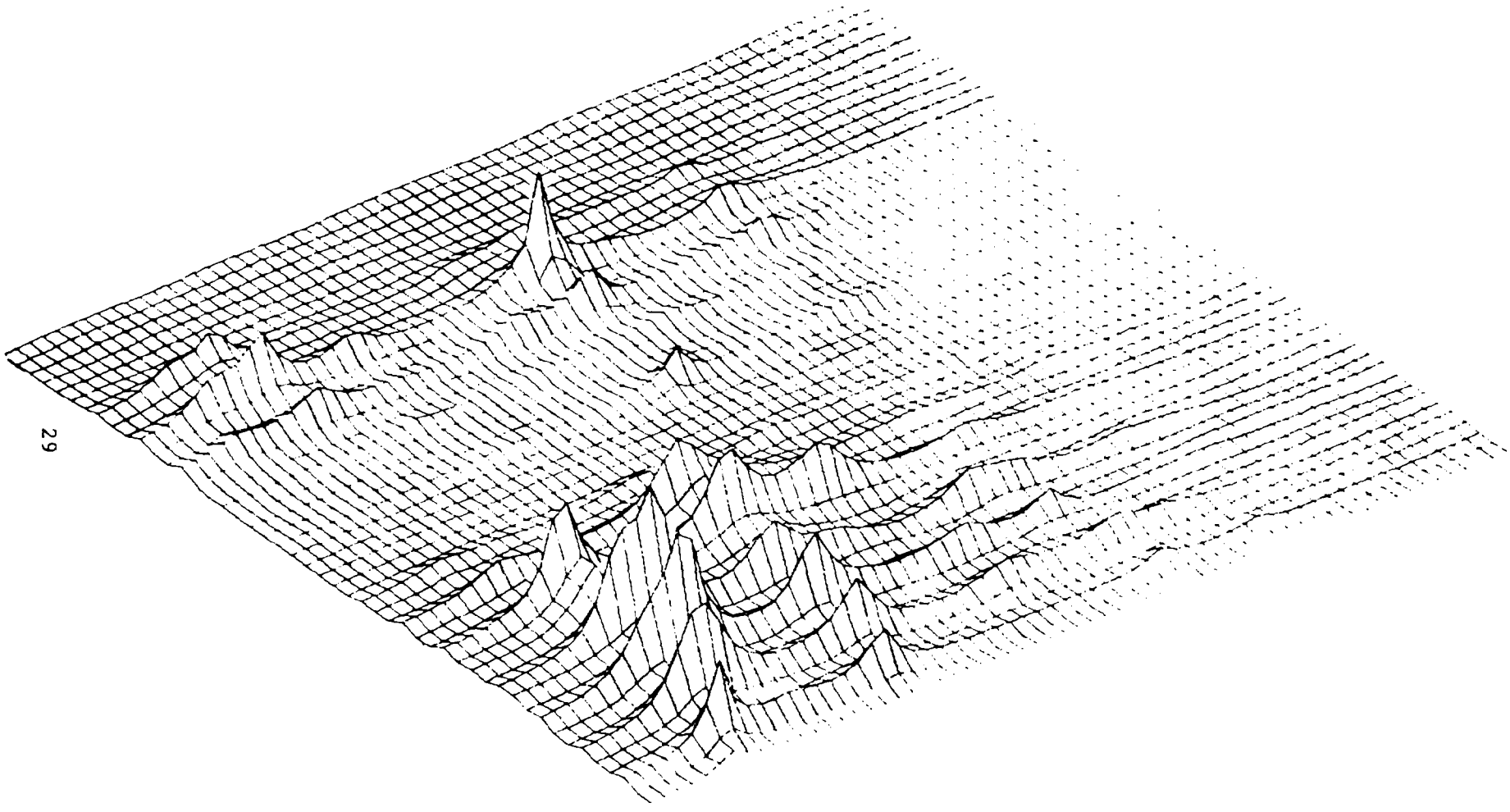


FIGURE 2.7 LPC SPECTROGRAM OF CVSD AT 16 KBPS (LL1)

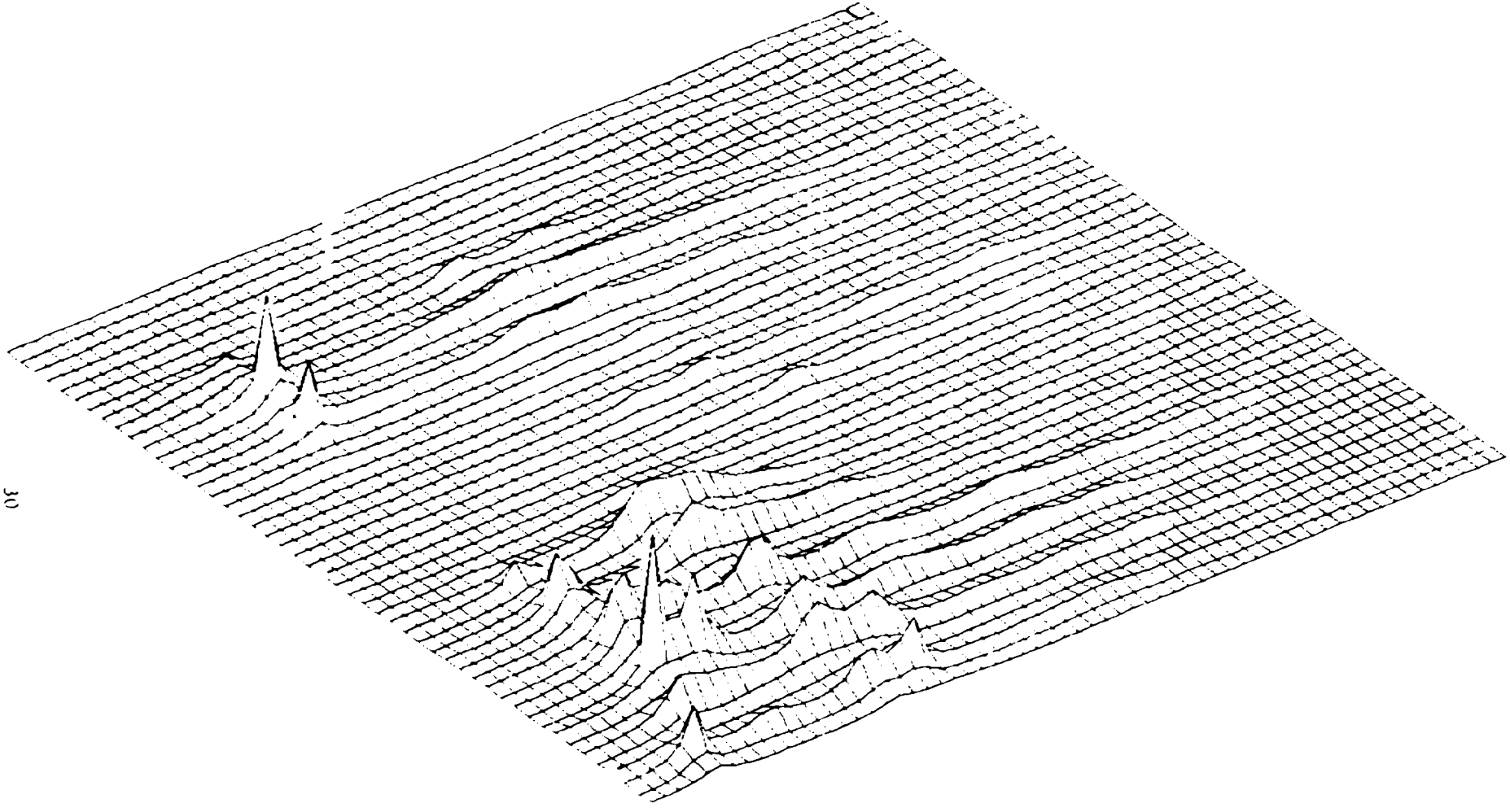


FIGURE 2.8 SPECTROGRAM OF HY2 AT 2.4 KBPS (LL1)

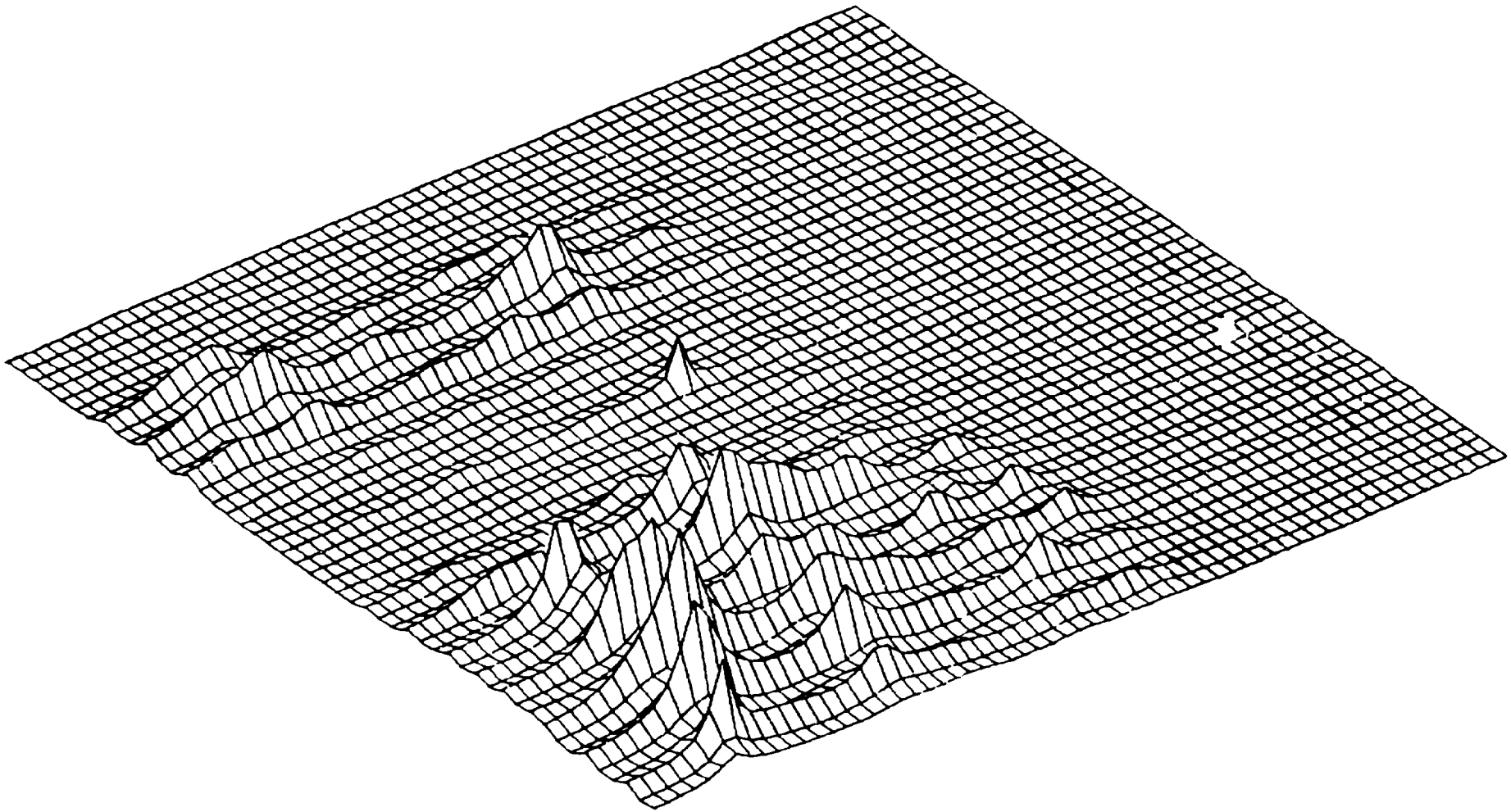


FIGURE 2.9 LPC SPECTROGRAM OF LONGBRAKE AT 2.4 KBPS (LL1)

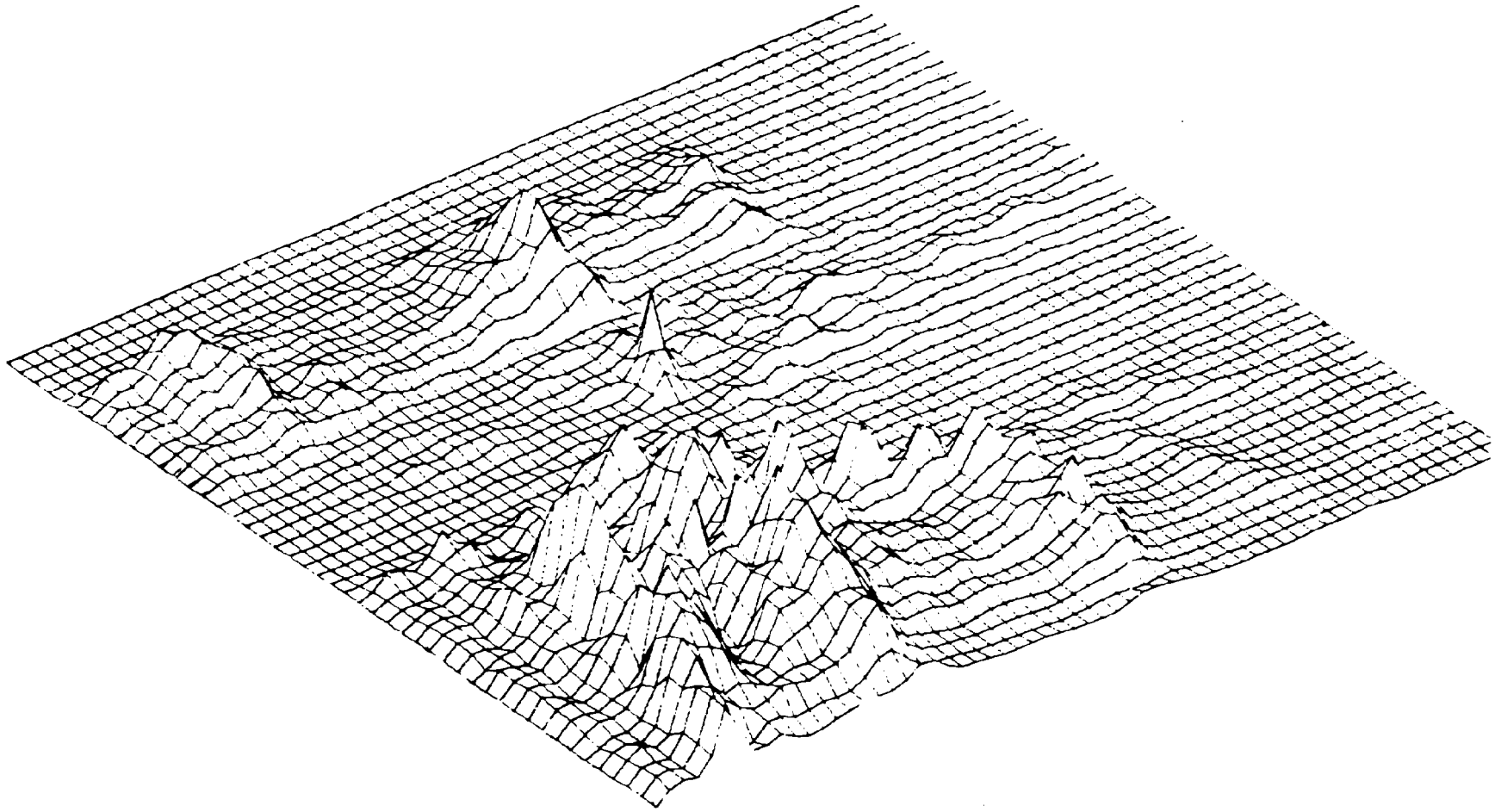


FIGURE 2.10 CEPSTRAL SPECTROGRAM OF HIANCHOR (LL1)

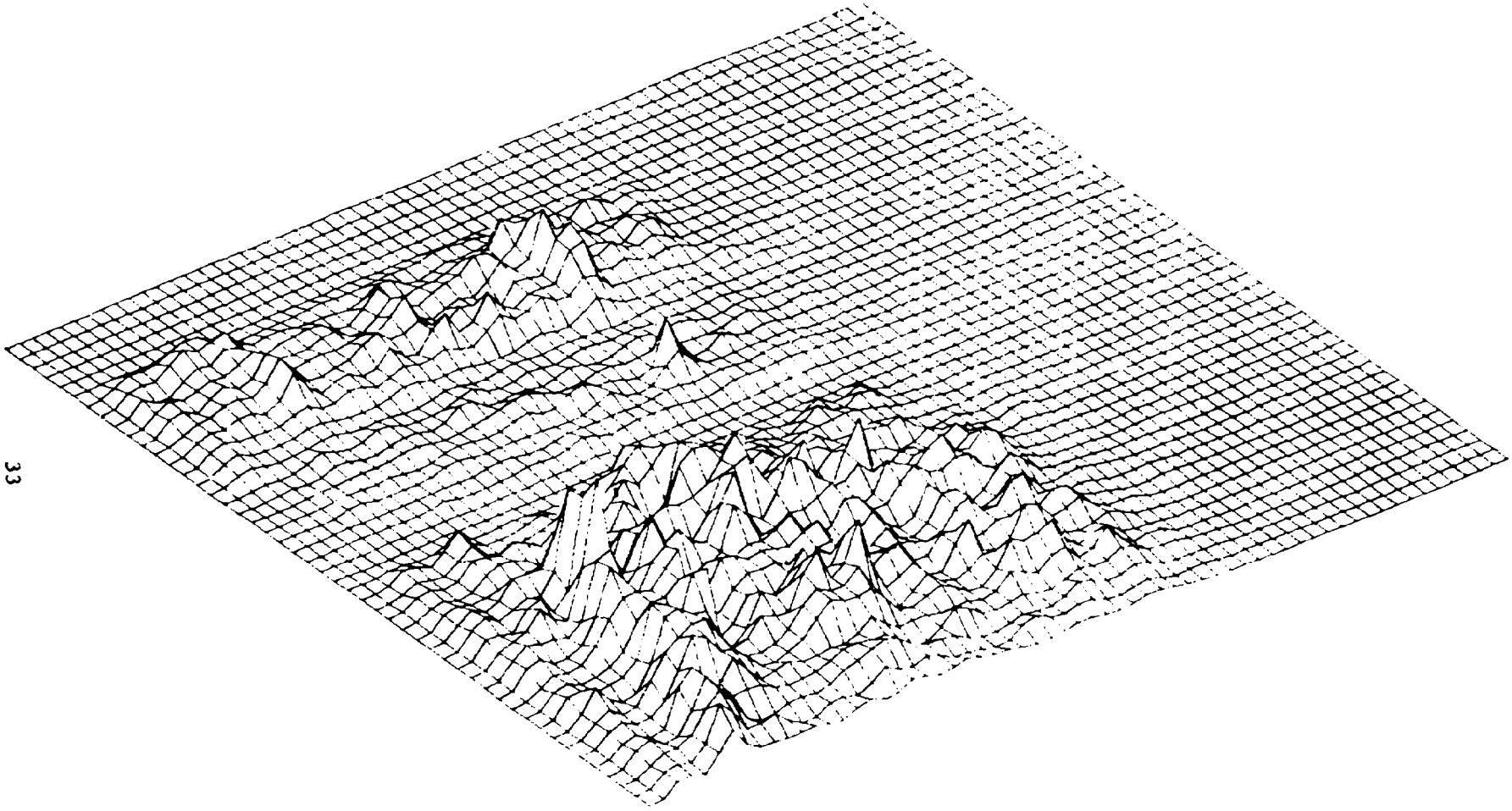


FIGURE 2.11 CEPSTRAL SPECTROGRAM OF CVSD AT 9.6 KBPS (LL1)

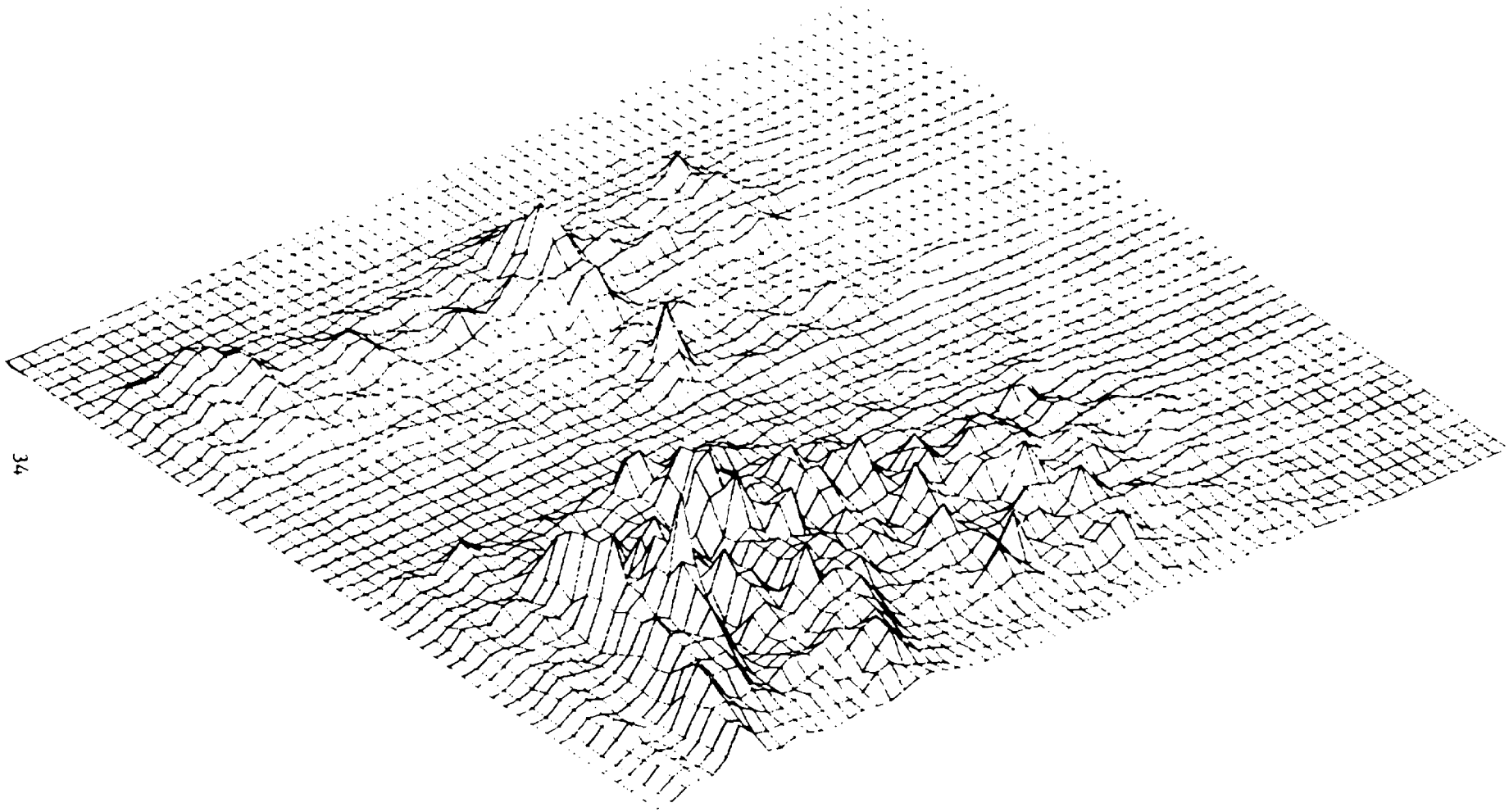


FIGURE 2.12 CEPSTRAL SPECTROGRAM OF CVSD AT 16 KBPS (LL1)

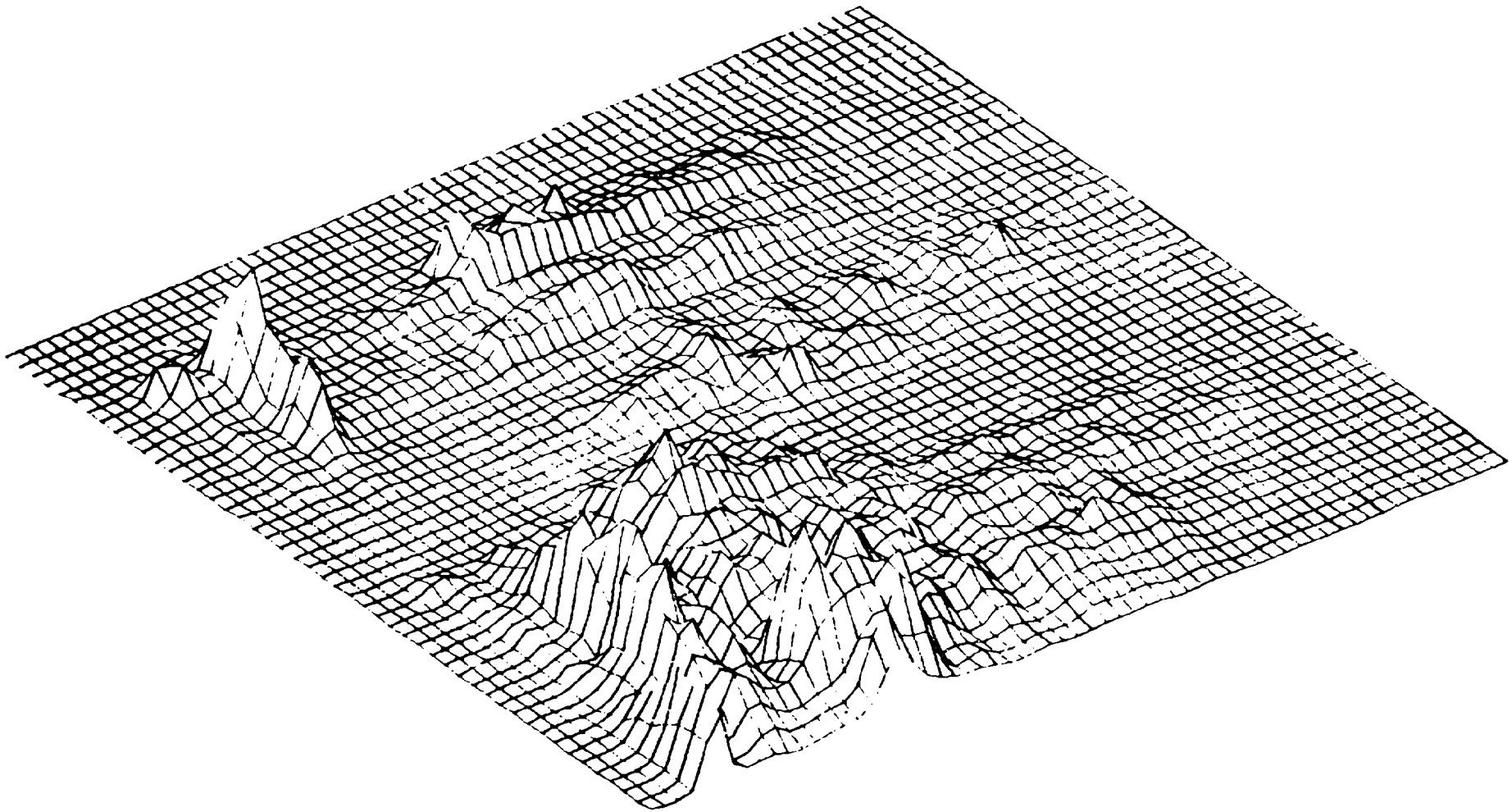


FIGURE 2.13 CEPSTRAL SPECTROGRAM OF HY2 AT 2.4 KBPS (LL1)

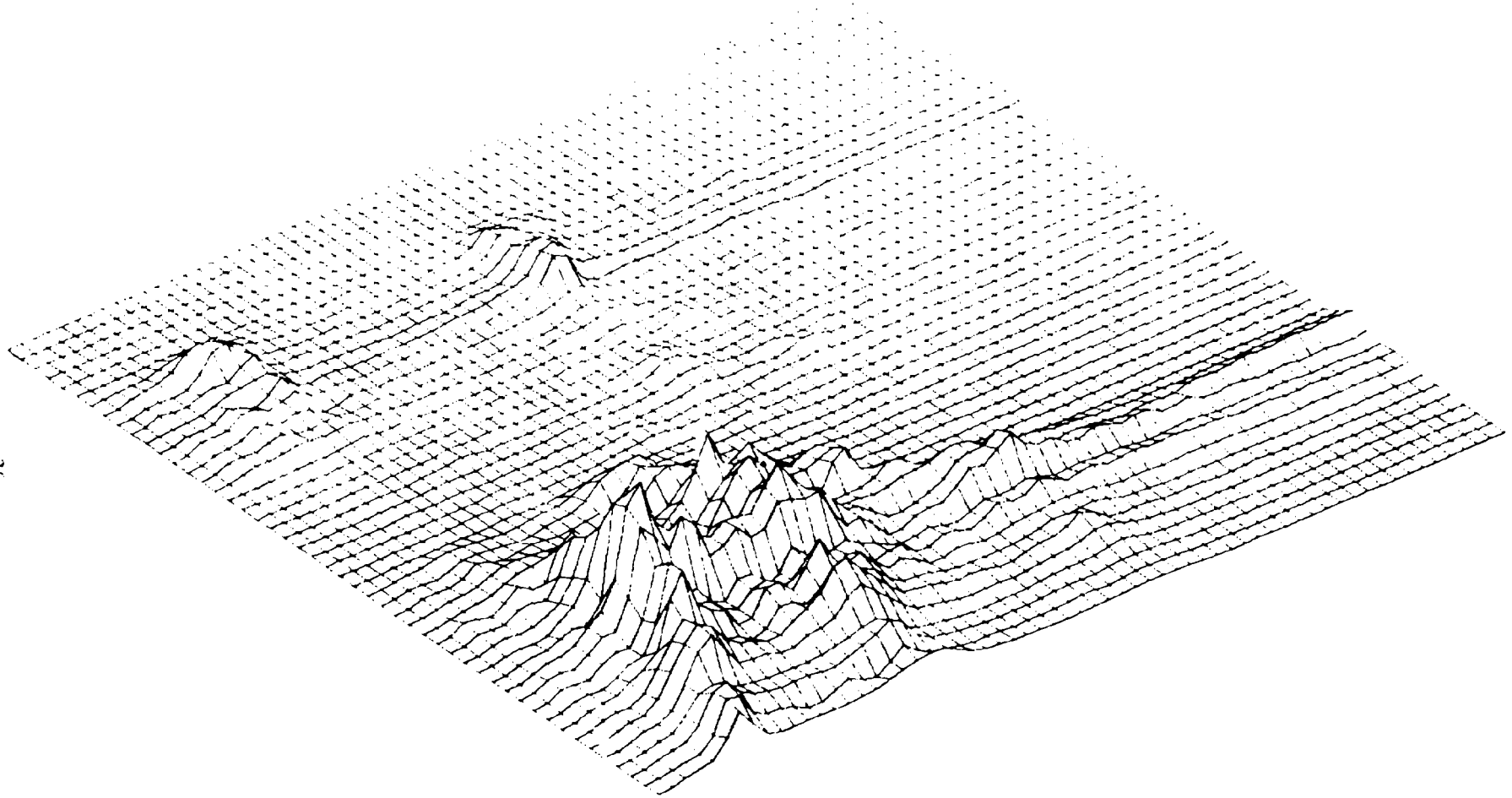


FIGURE 2.14 CEPSTRAL SPECTROGRAM OF LONGBRAKE AT 2.4 KBPS (LL1)

2.3.2 The Controlled Distortion Experiment

The purpose of the controlled distortion experiments was to test the candidate measures discussed in Section 2.2 as to their resolving power for measuring certain classes of distortions. In all cases, the "original" was taken to be the output of a 12 tap LPC synthesis program where the coefficients were unquantized and the pitch was extracted by hand. Two sets of signals were used. One set consisted of four synthetic vowels (/i/,/æ/,/u/ and /ɑ/), the other of two sentences, one spoken by a male speaker and one spoken by a female speaker. In all cases, five classes of distortions were applied: bandwidth distortion; frequency distortion; pitch distortion; low pass filtering distortion; and additive noise.

2.3.2.1 Bandwidth Distortion

Distortions in the bandwidth of formants is a common occurrence in vocoders. To test this type of distortion, the unit circle was effectively expanded by transforming each LPC coefficient by

$$a_i \rightarrow a_i(\alpha) \ . \qquad 2.19$$

In this experiment, the four values of α which were used were .99, .98, .97, and .95. The first two values introduced no perceivable distortion.

2.3.2.2 Frequency Distortion

The frequency distortion was carried out by up or down sampling the impulse response of the LPC synthesizer. Figure 2.15 shows the procedure. First, a FIR (256 point) approximation for the IIR impulse response was calculated. Then a zero padded interpolation was performed using a 1000 point Kaiser window designed linear phase low pass filter. The resulting modified impulse response was used to synthesize the

speech samples. Sampling ratios of 49-50, 50-49, 9-10, and 10-9 were used.

2.3.2.3 Pitch Distortion

Pitch distortion was applied by allowing the pitch period to systematically increase over the voiced regions. This results in pitch distortions which increased with time in each utterance. The rates at which the periods were allowed to vary was +1 sample every 10 voiced frames, +1 sample every 4 voiced frames, -1 sample every 10 voiced frames, and -1 sample every 4 voiced frames.

2.3.2.4 Low Pass Filter Distortion

Bandlimiting distortions are very common in speech communication systems, and hence worthy of study. The filters used were all 10^{th} order recursive digital eliptical filters with rejection bands at -60 DB. In all, four filters were used with cutoffs at 1.4 kHz, 1.8 kHz, 2.2 kHz, and 2.8 kHz.

2.3.2.5 Additive White Noise Distortion

White Gaussian noise was also added to the test signals. Four noise levels were used which resulted in signal to noise ratios of ~ 13 db, ~ 10 db, ~ 7 db, and ~ 3 db.

2.3.3 The Experimental Results

In all, six utterances, four vowels .768 seconds in length and two sentences 3.072 seconds in length, were used as originals. A total of four distortions for each of the five classes were applied to the six speech samples, giving 120 distorted samples. The purpose of the vowel distortion study was to measure the effects of each measure in a "micro" sense in order to compare resolving powers of the different measures. The purpose of the full sentence distortions was to measure the "macro"

behavior of each objective measure. In all cases, the total sentence metric was calculated from

$$D_p = \frac{\sum_{m=1}^{M} W'(m) d_{p,m}}{\sum_{m=1}^{M} W'(m)} \quad . \qquad 2.19$$

In this expression, D_p is the total distortion for the entire sentence set, $W'(m)$ is a weighting function, $d_{p,m}$ is the "d_p" measures defined in Section 2.2.2 at the m^{th} frame of the analysis, and M is the total number of analysis frames. $W'(m)$ was taken to be

$$W'(m) = 1, \qquad 2.20$$

and

$$W'(m) = G_m \quad , \qquad 2.21$$

where G_m is the LPC gain of the original sentence in the m^{th} frame. The LPC analyses were always done with a Hamming windowed, autocorrelation LPC with a frame interval of 256 samples and a window width of 256 samples. The gain weighting here was included to see how the overall outcome would be effected as a matter of academic interest. The hypothesis is that, since the vocalics contain a large portion of the information, and since the gain is always greater for vocalics, then a gain weighted measure might be more highly correlated with perceptual results. This experiment, clearly, gives no new information on this hypothesis, but it does show to what extent gain weighting changes the final objective quality estimate.

In all cases, D_p was taken to be the sum of M independent random

variables, all with the same standard deviation. The sample variance was calculated from

$$\hat{\sigma}_{D_p} = \sqrt{\sum_{m=1}^{M} \frac{(d_{p,m} - D_p)^2}{M-1}} \quad . \qquad 2.22$$

The random variable

$$t = \frac{D_p - \bar{D}_p}{\hat{\sigma}_{D_p}} \qquad 2.23$$

is t distributed (see Chapter 3) with zero mean and unit variance. A confidence interval for $\bar{D}_p$, the true mean for D_p, for a significance level α (α = .01 and .05) can be calculated from

$$D_p - U_{\alpha M}\hat{\sigma}_{D_p} < \bar{D}_p < D_p - L_{\alpha M}\hat{\sigma}_{D_p} \qquad 2.24$$

where $L_{\alpha M}$ and $U_{\alpha M}$ are the lower and upper significance limits for a t distributed random variable (μ = 0, σ = 1) for M points and probability α.

2.3.3.1 Results of the Vowel Tests

The results of the vowel tests for frequency distortion and bandwidth distortion are compiled in Table 2.3, the results for low pass filtering distortion and noise distortion are given in Table 2.4, and the results for pitch distortion are given in Table 2.5.

Several points should be made about these results. First, all of the tests seem to perform relatively well on the two frequency distortions, with all tests able to resolve the distortions at least

SPECIAL DISTORTION MEASURES		BANDWIDTH DISTORTIONS α				FREQUENCY SHIFT DISTORTIONS SHIFT RATIOS			
		.99	.98	.97	.95	50/49	49/50	10/9	9/10
D_1 LOG LPC	AV.	.076	.13	.22	.37	.08	.07	.91	.83
	C.I.	.03	.04	.06	.12	.03	.03	.11	.10
D_2 LOG LPC	AV.	.081	.21	.24	.46	.11	.10	1.2	.90
	C.I.	.03	.05	.04	.12	.04	.02	.12	.10
D_4 LOG LPC	AV.	.12	.26	.33	.61	.13	.15	1.6	1.3
	C.I.	.05	.06	.09	.17	.05	.05	.14	.12
D_2 LINEAR LPC	AV.	1280	1541	3021	4077	2041	2112	4510	4910
	C.I.	825	1051	1121	1642	914	921	2013	2412
D_1 CEPSTRUM	AV.	.088	.22	.25	.42	.14	.12	1.3	.91
	C.I.	.03	.05	.06	.13	.03	.03	.11	.11
D_2 PARCOR	AV.	1.1	1.6	1.8	2.3	1.5	1.3	3.2	2.1
	C.I.	.06	.05	.07	.08	.04	.02	1.2	.09
D_2 FEEDBACK	AV.	113	191	215	421	104	127	411	402
	C.I.	61	75	112	181	55	67	172	101
D_2 AREA	AV.	1.1	2.2	3.1	5.7	1.4	1.2	3.7	3.2
	C.I.	0.2	0.2	0.4	1.1	.31	.32	.62	.59
D_2 POLE LOCATION	AV.	2.3	2.7	2.9	4.1	2.1	1.9	4.2	3.8
	C.I.	.93	1.6	1.9	2.2	.91	.80	2.1	2.3

AV. = Average C.I. = Confidence Interval (.05 Level)

Table 2.3 Results of the Bandwidth Distortions and Frequency Shift Distortions on Vowels. All Confidence Intervals are at the .05 Level.

SPECIAL DISTORTION MEASURES		BANDLIMIT DISTORTION				NOISE DISTORTION			
		BANDLIMIT				S/N			
		2.8	2.2	1.8	1.4	13	10	7	3
D_1 LOG LPC	AV.	7.3	12.1	14.6	16.2	1.7	2.8	5.0	7.8
	C.I.	1.1	2.4	2.8	3.5	.22	.62	.97	1.81
D_2 LOG LPC	AV.	8.1	13.3	15.6	17.5	1.9	3.2	5.2	8.6
	C.I.	1.2	2.3	3.1	3.6	.31	.82	1.4	2.6
D_4 LOG LPC	AV.	9.4	14.4	16.7	18.2	2.4	3.6	5.6	10.1
	C.I.	1.4	2.5	3.5	3.7	.40	1.02	1.05	1.19
D_2 LINEAR LPC	AV.	6851	7175	8283	9143	5431	5941	6643	7141
	C.I.	855	991	1097	1211	2413	2712	3143	4127
D_2 CEPSTRUM	AV.	8.8	14.1	16.0	18.1	1.6	3.1	5.2	8.8
	C.I.	1.3	2.2	3.3	3.6	.33	.91	1.3	2.7
D_2 PARCOR	AV.	5.2	5.5	5.9	6.3	3.1	3.6	4.3	4.6
	C.I.	1.1	1.3	1.2	1.6	.81	.80	.93	.92
D_2 FEEDBACK	AV.	827	955	1010	1210	621	751	827	921
	C.I.	310	341	381	425	125	281	317	397
D_2 AREA	AV.	5.3	5.9	6.6	6.9	2.8	2.9	3.1	3.3
	C.I.	.34	.41	.55	.57	.21	.35	.44	.89
D_2 POLE LOCATION	AV.	6.6	6.7	6.7	6.9	4.1	4.4	4.9	5.2
	C.I.	3.4	3.3	3.3	3.6	2.2	2.1	2.7	2.6

AV. = Average C.I. = Confidence Interval (.05 Level)

Table 2.4 Results of the Bandlimit Distortion and Additive Noise Distortion on Vowels. All Confidence Intervals Are at the .05 Level.

the .05 level. This point is also illustrated in Figures 2.15 and 2.16, which show the time behavior of the d_2 log LPC measure for the frequency and bandwidth distortion. As judged by their confidence intervals, the log LPC measures are the best, while the pole position and feedback coefficients are the worst for these two frequency distortions. Second, note that, for low pass filter distortion (Table 2.4), the results are qualitatively the same as those above. But also note that quantitatively they are very different, giving much greater spectral distances than the bandwidth and frequency shift distortions. This can also be seen in Figure 2.17. This brings up an important, if obvious, point. That is that low pass filtering distortion swamps the more subtle forms of frequency distortion. Hence, some bandwidth decision and control is necessary in these objective tests if the more subtle distortions are to be measured.

The noise results show some resolving power for the various noise levels, but a general loss of resolution when compared to the frequency and bandwidth results. Stated simply, this type of distortion is not measured well by spectral distance measures, and hence requires a large sample of speech to detect it properly.

The results of the pitch variation studies presented in Table 2.5 show that essentially no spectral distance measure can detect pitch errors with the number of samples used in this experiment. This, of course, was an expected result, and was the reason that the special pitch tests were included.

The cepstral pitch measure described in Section 2.2.2.2 was applied to the four pitch distortions using each of the four smoothing window functions shown in Figure 2.17.

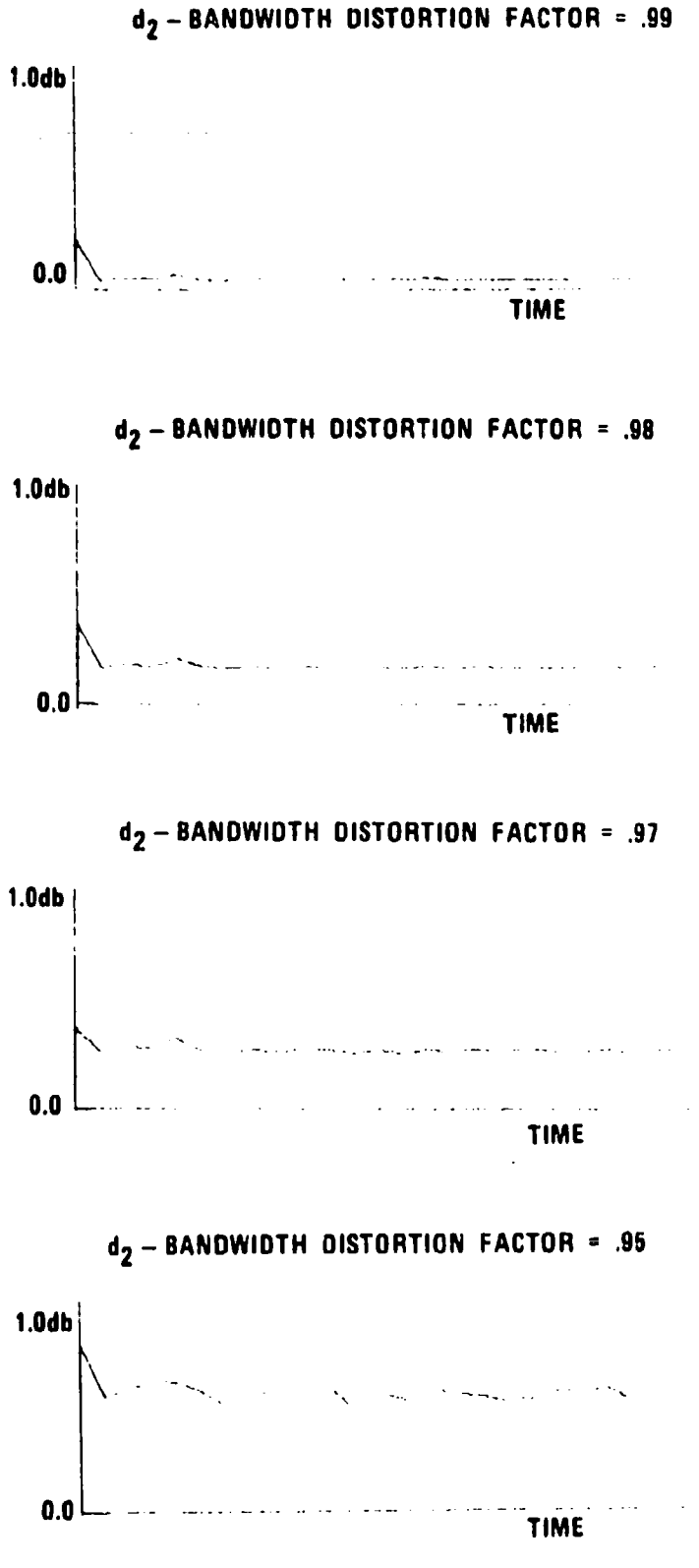


FIGURE 2.15 PLOTS OF d_2 LOG LPC SPECTRAL DISTANCE MEASURES FOR THE SYNTHETIC VOWEL FOR VARIOUS BANDWIDTH DISTORTION FACTORS. THE DISTORTION IS FORMED FROM $a_1 \sigma^1 a_1$ WHERE σ IS THE BANDWIDTH DISTORTION FACTOR.

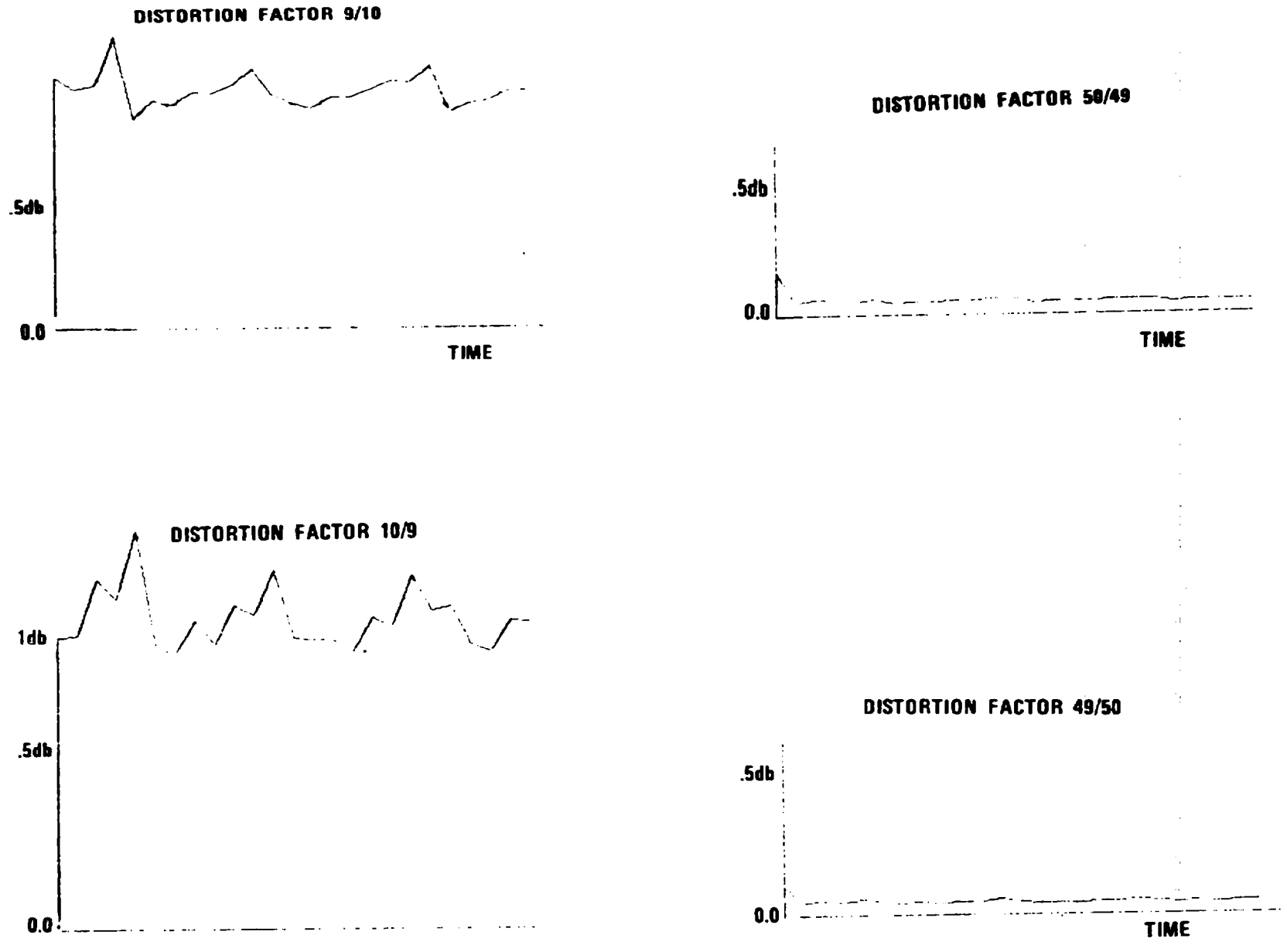


FIGURE 2.16 PLOTS OF d_2 LOG LPC SPECTRAL DISTANCE MEASURES FOR THE SYNTHETIC VOWEL /æ/ FOR VARIOUS FREQUENCY SHIFT DISTORTION RATIOS.

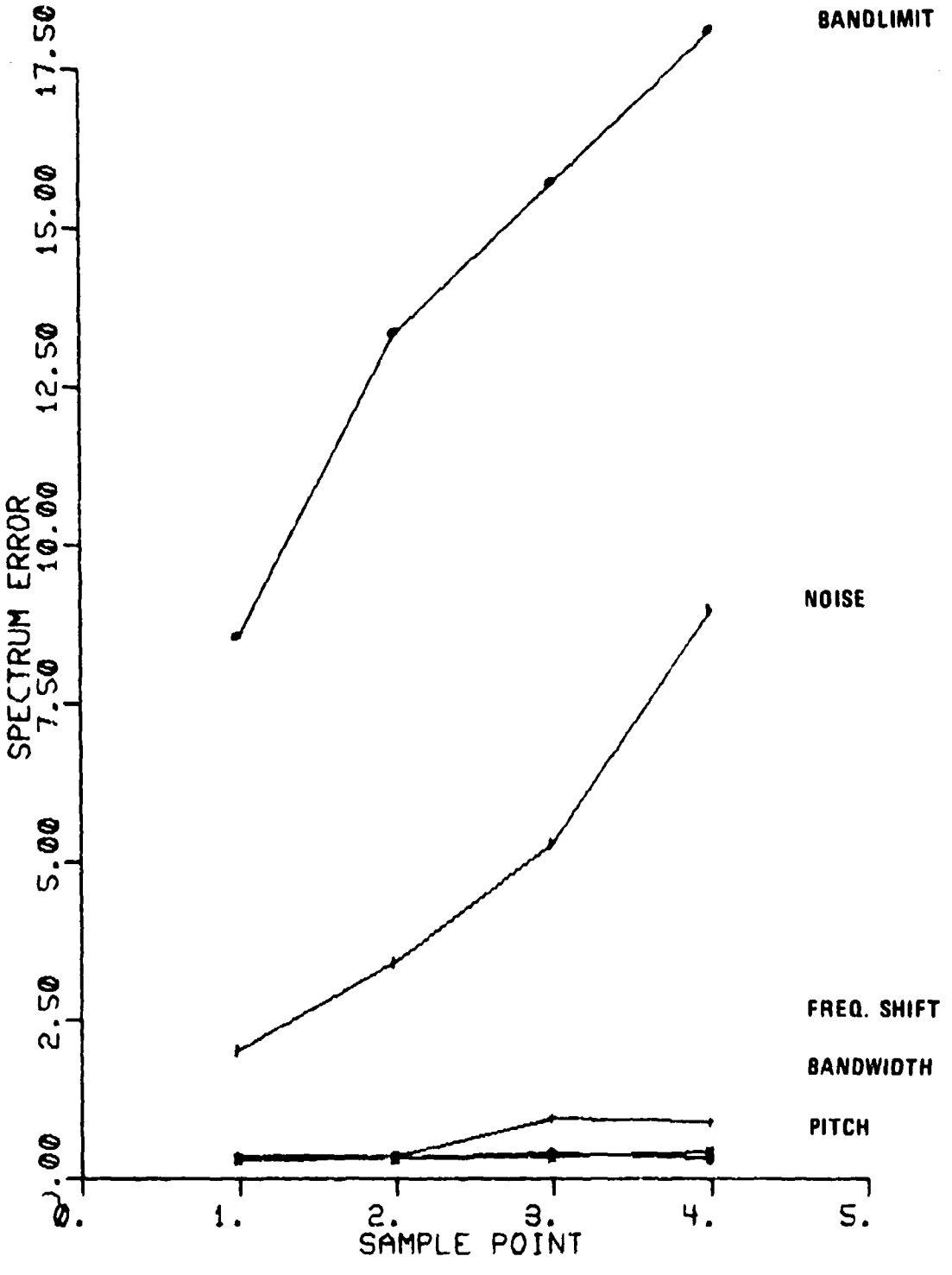


FIGURE 2.17(a) PLOTS OF THE d_2 LOG LPC SPECTRAL DISTANCE MEASURE ON VOWELS FOR THE VARIOUS DISTORTIONS USED IN THIS STUDY.

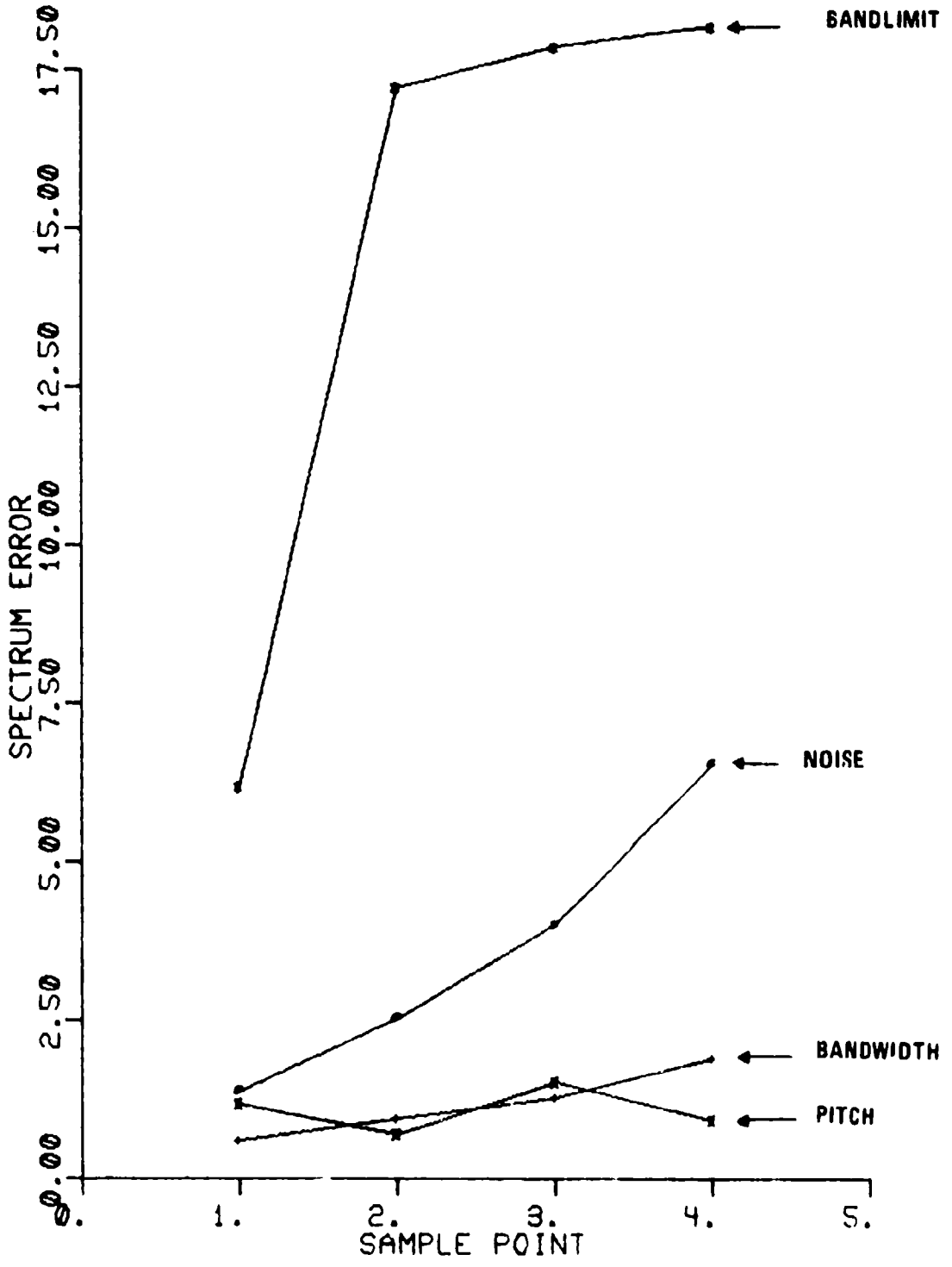


FIGURE 2.17(b) PLOTS OF THE d_2 LOG LPC SPECTRAL DISTANCE MEASURES ON SENTENCES FOR THE VARIOUS DISTORTIONS USED IN THIS STUDY.

SPECTRAL DISTORTION MEASURES		PITCH DISTORTION			
		10,1	10,-1	4,1	4,-1
D_1 LOG LPC	AV.	.071	.064	.073	.072
	C.I.	.03	.03	.04	.03
D_2 LOG LPC	AV.	.079	.081	.076	.078
	C.I.	.03	.03	.03	.03
D_4 LOG LPC	AV.	.09	.092	.084	.092
	C.I.	.04	.05	.04	.04
D_2 LINEAR LPC	AV.	821	871	888	841
	C.I.	640	510	530	511
D_1 CEPSTRUM	AV.	.82	.86	.84	.81
	C.I.	.03	.03	.04	.03
D_2 PARCOR	AV.	.91	.84	.88	.86
	C.I.	.06	.05	.06	.05
D_2 FEEDBACK	AV.	87	88	83	89
	C.I.	48	51	55	46
D_2 AREA	AV.	.91	.96	.81	.86
	C.I.	.21	.23	.20	.19
D_2 POLE LOCATION	AV.	2.1	2.0	2.2	2.3
	C.I.	1.10	1.02	1.05	.98

AV. = Average

C.I. = Confidence Interval (.05 Level)

Table 2.5 Results of the Pitch Distortions on Vowels. Note that the Distortions are Low, and Increase Distortions Cause No Increase in the Measures.

Since this was a time varying distortion, then the statistical analysis used in the spectral distance tests is inappropriate. Figures 2.18-2.21 show the results for the four windows. The basic result here is that this measure forms a high resolution measure of pitch errors. For short windows, the measure detects very small errors, but saturates quickly, hence reporting the same result for all errors. Longer windows do a better qualification of the pitch errors, but do not pick up small errors well. Probably, since most of the computation in this measure is in the cepstrum calculation, a reasonable solution would be to use several windows to better quantify the results.

2.3.3.2 Results of the Sentence Tests

The results of the sentence tests are tabulated in Table 2.6, 2.7, and 2.8. Qualitatively, these results pretty well mirror the results of the vowel tests. Quantitatively, however, the confidence intervals are uniformly larger. The general result here, therefore, is that larger sample sizes are necessary when dealing with real sentences.

An important result of the sentence tests can be seen from a comparison of the gain weighted measures to the non gain weighted measures, as shown in Table 2.9. In nearly every case, the gain weighting causes the measure to decrease. This means the measure is being inflated by the low power unvoiced regions which are perceptually less important than the high vocalic regions. This means that gain weighting probably will give better subjective correlation.

2.4 The PARM Correlation Study

As was stated in the introduction, the PARM subjective quality data base offers a good chance to study the correlation between the objective measures under consideration and the isometric subjective

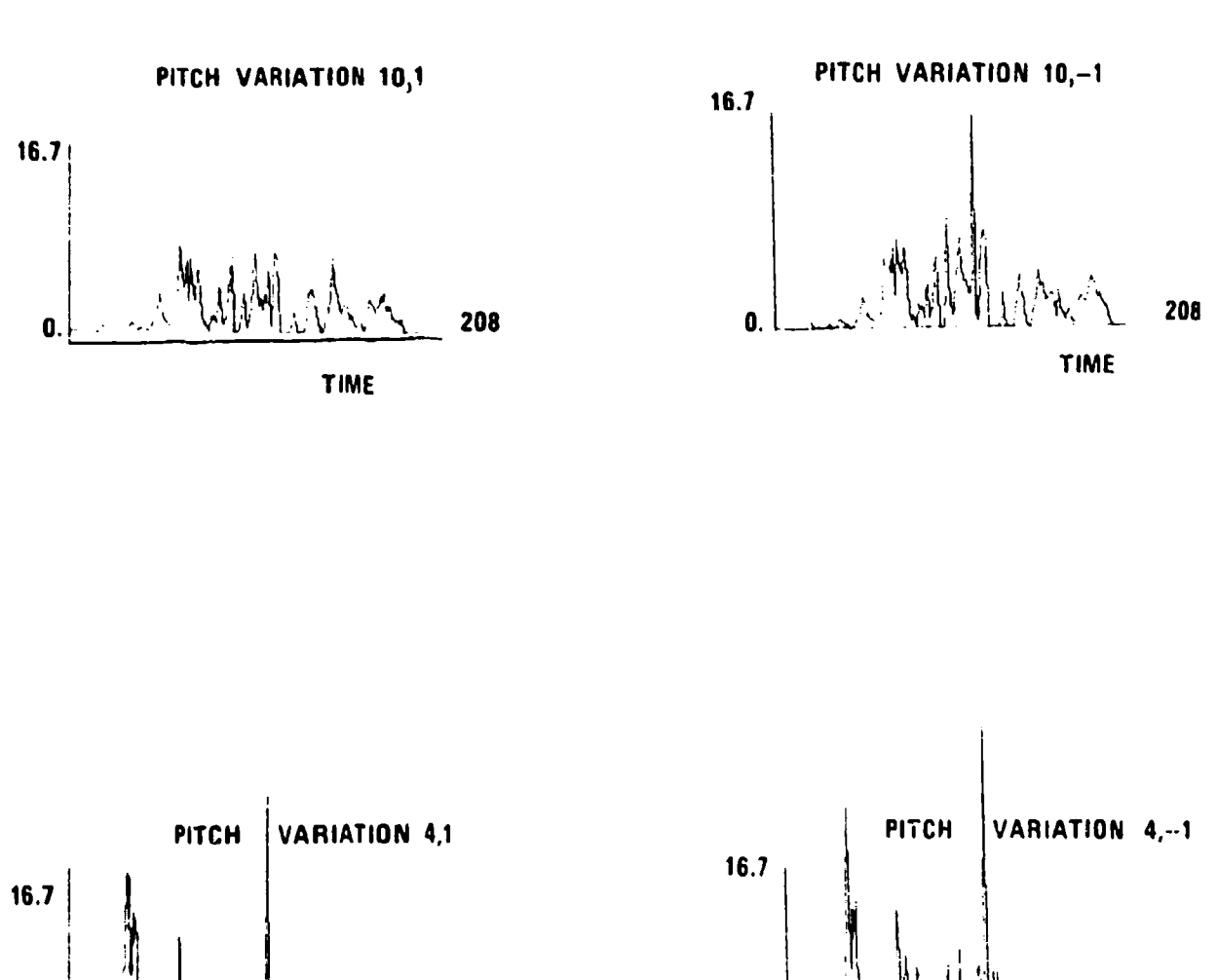


FIGURE 2.18 CEPSTRAL PITCH METRIC AS A FUNCTION OF TIME FOR FOUR DIFFERENT PITCH DISTORTIONS FOR WINDOW NO. 1 (FIGURE 2.3). WINDOW LENGTH = 1.

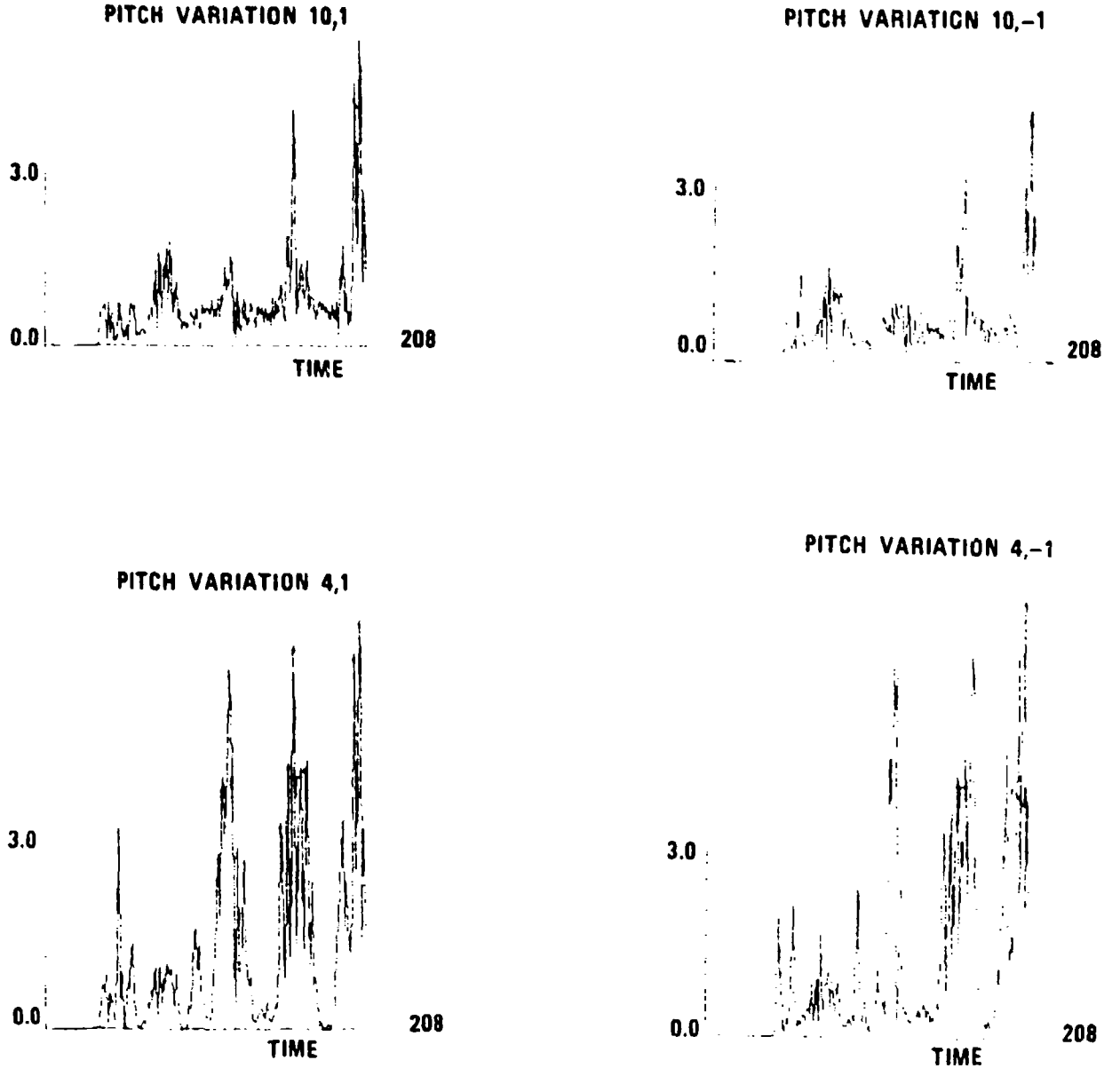


FIGURE 2.19 CEPSTRAL PITCH METRIC AS A FUNCTION OF TIME FOR FOUR DIFFERENT PITCH DISTORTIONS FOR WINDOW NO. 2 (FIGURE 2.3) WINDOW LENGTH = 4.

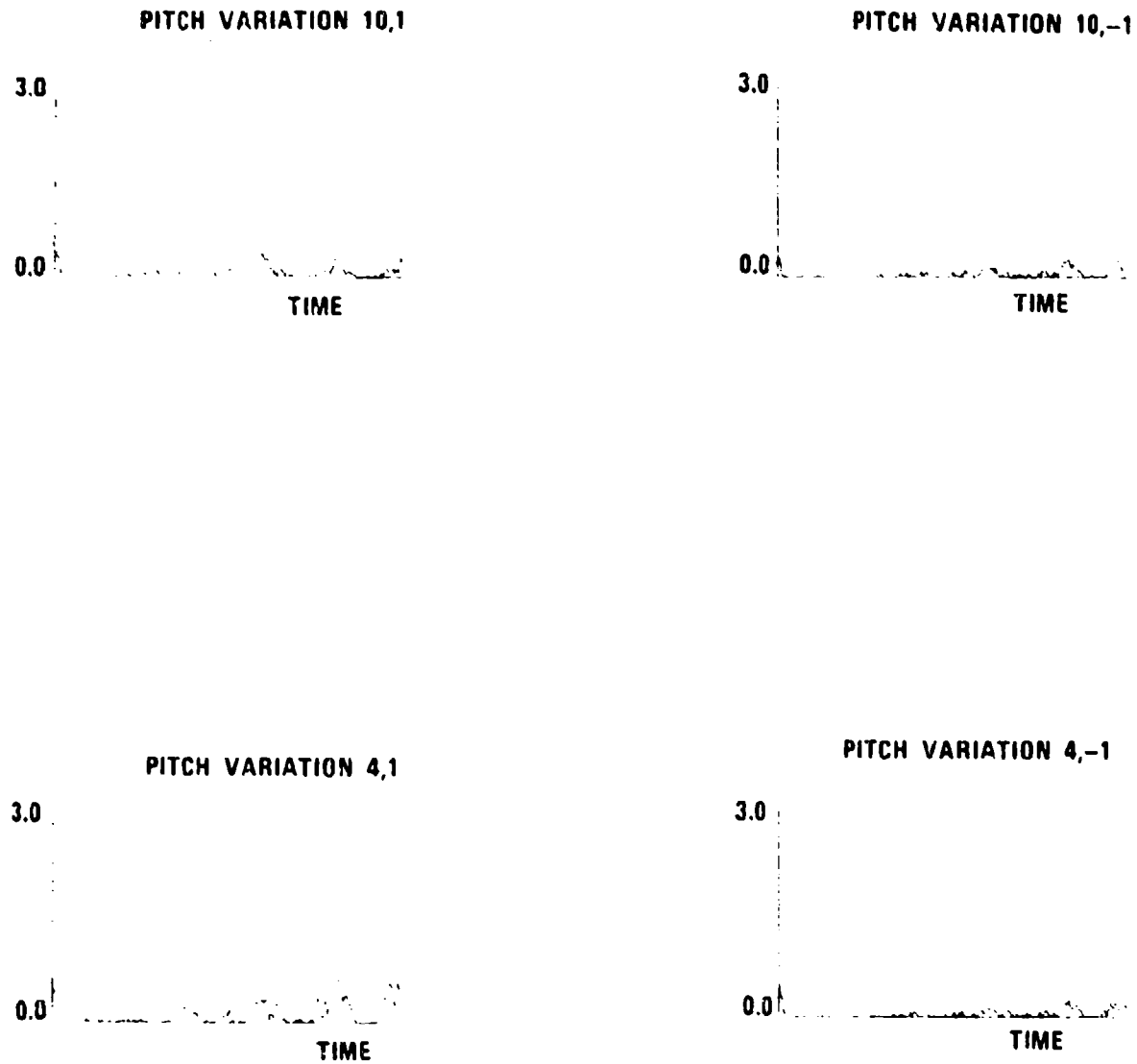


FIGURE 2.20 CEPSTRAL PITCH METRIC AS A FUNCTION OF TIME FOR FOUR DIFFERENT PITCH DISTORTIONS FOR WINDOW NO. 3 (FIGURE 2.3). WINDOW LENGTH = 10.

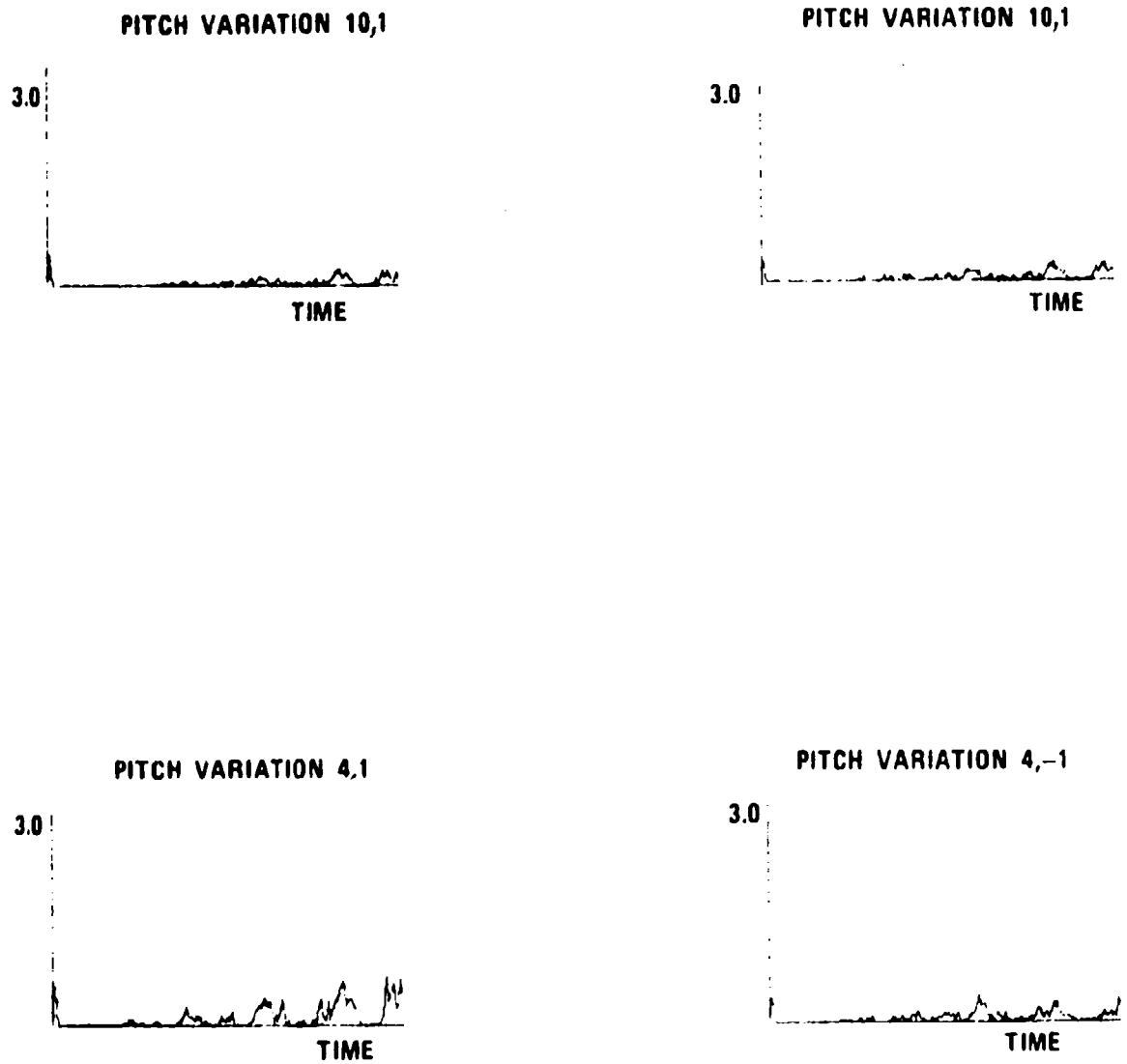


FIGURE 2.21 CEPSTRAL PITCH METRIC AS A FUNCTION OF TIME FOR FOUR DIFFERENT DISTORTIONS FOR WINDOW NO. 4 (FIGURE 2.3). WINDOW LENGTH = 10.

SPECTRAL DISTORTION MEASURES		BANDWIDTH DISTORTIONS				FREQUENCY SHIFT DISTORTIONS			
		α				SHIFT RATIOS			
		.99	.98	.97	.95	50/49	49/50	10/9	9/10
D_1 LOG LPC	AV.	.54	.88	1.2	1.6	.61	.58	1.7	1.9
	C.I.	.13	.13	.16	.22	.13	.12	.19	.24
D_2 LOG LPC	AV.	.62	.94	1.56	1.9	.71	.68	2.4	2.2
	C.I.	.12	.14	.17	.23	.14	.13	.27	.28
D_4 LOG LPC	AV.	.83	1.21	1.8	2.2	.94	1.02	3.1	3.4
	C.I.	.13	.16	.19	.24	.18	.16	.29	.29
D_2 LINEAR LPC	AV.	2910	3816	4715	6144	3415	2916	6913	6314
	C.I.	2010	2415	3103	3310	2413	1918	3412	3321
D_1 CEPSTRUM	AV.	.75	1.05	1.60	2.0	.82	.77	1.96	2.1
	C.I.	.14	.14	.19	.23	.15	.16	.3	.29
D_2 PARCOR	AV.	2.4	2.9	2.9	4.1	1.9	1.8	4.1	3.2
	C.I.	1.6	1.5	1.7	2.2	1.2	1.0	2.1	1.8
D_2 FEEDBACK	AV.	420	461	520	850	480	455	1023	981
		225	251	312	515	310	295	612	580
D_2 AREA	AV.	3.4	3.9	5.9	8.2	3.3	3.5	8.1	8.1
	C.I.	1.2	1.3	2.4	4.2	1.4	1.1	3.4	4.1
D_2 POLE LOCATION	AV.	4.6	4.9	5.4	6.3	4.8	4.6	6.8	6.3
	C.I.	2.4	3.1	4.1	4.8	3.1	2.8	4.4	4.2

AV. = Average C.I. = Confidence Intervals

Table 2.6 Results of the Bandwidth Distortions and Frequency Shift Distortions on Sentences. All Confidence Intervals are at the .05 Levels.

SPECTRAL DISTORTION MEASURES		BANDLIMIT DISTORTION				NOISE DISTORTION			
		BANDLIMIT				S/N			
		2.8	2.2	1.8	1.4	13	10	7	3
D_1 LOG LPC	AV.	7.5	15.4	16.8	17.2	1.1	2.1	3.8	5.7
	C.I.	2.7	5.8	5.7	9.6	.51	1.2	1.7	2.6
D_2 LOG LPC	AV.	6.1	16.3	16.9	17.5	1.2	2.4	4.1	6.6
	C.I.	1.3	7.2	7.1	9.2	.62	1.4	2.6	3.8
D_4 LOG LPC	AV.	8.4	16.2	16.8	17.5	1.6	2.9	4.7	6.3
	C.I.	1.5	6.8	7.5	8.2	.77	1.31	2.6	3.5
D_2 LINEAR LPC	AV.	8142	9317	9581	9721	4213	5176	6612	7123
	C.I.	2014	2713	2312	3140	2913	2310	3412	3731
D_1 CEPSTRUM	AV.	5.4	8.3	12.4	16.3	1.4	2.2	3.6	5.9
	C.I.	1.3	2.2	3.1	4.4	.52	1.3	2.2	2.9
D_2 PARCOR	AV.	7.1	8.3	8.9	9.2	6.2	6.7	7.7	9.2
	C.I.	3.6	3.9	4.7	5.3	4.4	4.5	5.3	6.1
D_2 FEEDBACK	AV.	1013	1314	1517	1712	823	941	1021	1313
	C.I.	712	692	851	1003	512	590	610	713
D_2 AREA	AV.	6.7	7.3	8.2	8.8	4.2	4.4	4.7	5.1
	C.I.	1.3	1.9	2.3	2.6	2.1	2.2	2.8	3.3
D_2 POLE LOCATION	AV.	7.2	7.7	7.5	7.8	6.3	6.7	6.8	7.2
	C.I.	4.4	4.7	3.9	4.6	3.1	3.6	3.2	4.1

AV. = Average C.I. = Confidence Interval (.05)

Table 2.7 Results of the Bandlimit Distortions and Additive Noise Distortion on Sentences. All Confidence Intervals are at the .05 Significance Level.

SPECTRAL DISTORTION MEASURES		PITCH DISTORTIONS			
		10,1	10,-1	4,1	4,-1
D_1 LOG LPC	AV.	1.0	1.1	.90	.97
	C.I.	.12	.31	.22	.24
D_2 LOG LPC	AV.	1.2	.63	1.5	.94
	C.I.	.25	.11	.09	.10
D_4 LOG LPC	AV.	1.4	1.2	1.8	1.7
	C.I.	.13	.15	.21	.19
D_2 LINEAR LPC	AV.	1041	981	1101	1315
	C.I.	512	412	520	640
D_1 CEPSTRUM	AV.	1.3	1.4	1.5	1.4
	C.I.	.04	.02	.03	.03
D_2 PARCOR	AV.	2.0	1.9	2.3	2.2
	C.I.	.92	.82	1.1	1.4
D_2 FEEDBACK	AV.	310	412	391	360
	C.I.	240	270	210	170
D_2 AREA	AV.	2.6	2.4	2.8	2.7
	C.I.	.62	.51	.83	.84
D_2 POLE LOCATION	AV.	3.8	3.9	4.2	4.0
	C.I.	1.8	1.9	2.4	2.6

AV. = Average

C.I. = Confidence Interval (.05)

Table 2.8 Results of the Pitch Distortion Study on Vowels. All Confidence Intervals are at the .05 Significance Level.

DISTORTION	NON-GAIN WEIGHTED	GAIN WEIGHTED
Bandwidth .99	.62	.38
Bandwidth .98	.94	.67
Bandwidth .97	1.56	1.64
Bandwidth .95	1.9	1.51
Frequency Shift 50/49	.71	.37
Frequency Shift 49/50	.68	.37
Frequency Shift 10/9	2.4	1.92
Frequency Shift 9/10	2.2	2.12
Bandlimit 2.8 kHz	6.1	4.3
Bandlimit 2.2 kHz	16.3	12.4
Bandlimit 1.8 kHz	16.9	14.7
Bandlimit 1.4 kHz	17.5	16.8
Noise 13 db	1.2	.82
Noise 10 db	2.4	1.81
Noise 7 db	4.1	3.6
Noise 3 dB	6.6	5.4

Table 2.9 Comparison of Gain Weighted D_2 Log LPC Spectral Metrics to Non-Gain Weighted D_2 Log LPC Spectral Metrics.

results available from the PARM. Since many of the objective measures under study are computationally intensive, the computer time limited the total number of speech digitization systems which could be used as part of the study. In all, eight systems were studied, as shown in Table 2.10. These systems were chosen to (1) represent a cross-section of speech digitization techniques, including waveform coders (CVSD), LPC's, channel vocoders, and APC's, and (2) these systems overlapped with the systems used in the development of a parametric quality test, called the "QUART" Test [2.24]. This allows some minimal correlation studies between the objective quality measures produced here and the parametric results available from the QUART test.

2.4.1 The PARM Data Base

The PARM data base arrived at Georgia Tech as fourteen boxes of cards, with control cards for processing under an IBM operating system. Since correlation studies require many accesses of the data base, and since the accesses are random, a linear data base such as that represented by the cards is unacceptable. An acceptable data base organization must (1) be stored in numeric (two's complement) form rather than character form, and (2) must be accessable by some coding scheme which does not require the linear searching of the disk based data. To do this, the system of Figure 2.22 was developed. In this system, a "MAIN DATA FILE" was organized in which each set of responses for each subject is allocated a direct accessable block of 64 sixteen bit words, 60 for the subject's responses and four for a label. To go with this main file, four "POINTER FILES" were developed. The first pointer file, the "PARM IDENTITY FILE." as an entry for each PARM giving basic PARM data, such as systems involved, speakers involved, and pointer to the main data file. The second pointer file, the "SPEAKER FILE," has

HI ANCHOR
1. CVSD - 32-0%
2. CVSD - 16-0%
3. CVSD - 9.6-0%
4. LPC - 4.8-0% (Lincoln Labs)
5. LPC - 3.6-0% (Lincoln Labs)
6. LPC - 2.4-0% (Lincoln Labs)
7. APC - 0%
8. PARKHILL - 20 db S/N
9. HY2 - 2.4-0%

Table 2.10 Systems Used in the PARM Correlation Study.

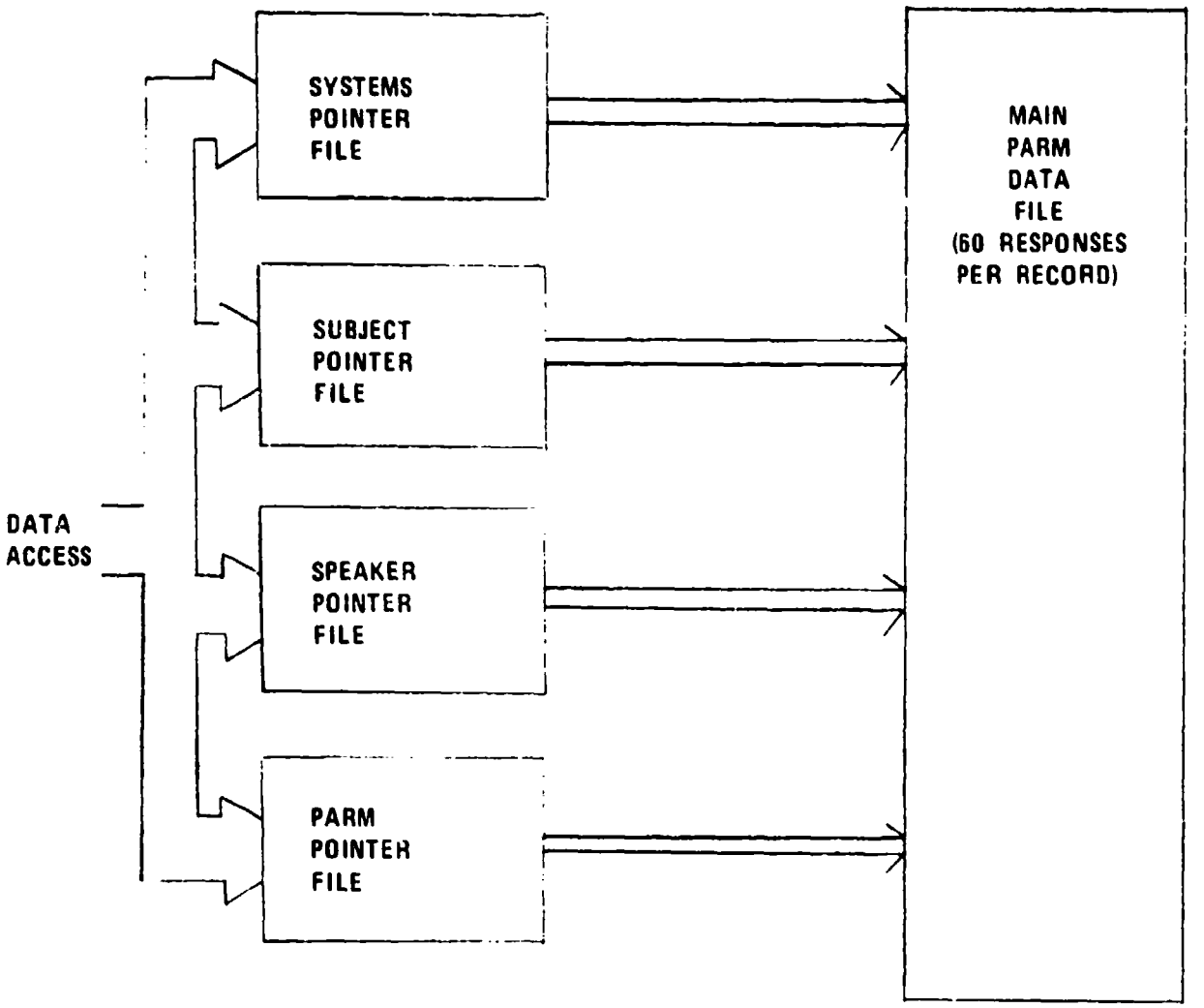


FIGURE 2.22 LAYOUT OF PARM ACCESS DATA USED AS PART OF THIS STUDY. EACH BOX REPRESENTS A DISK FILE. THE DATA IS PRESORTED IN THE DATA FILES TO ALLOW EASY ACCESS OF THE PARM DATA SETS.

information for each speaker as to where each PARM involving that speaker is located. The third file, the "SUBJECT FILE" contains a list, by subject, of where each of that subject's responses is located. The last pointer file, the "SYSTEM FILE" contains, for each system, the location of all that system's subjective data.

The idea behind this organization is that, by presorting on the information of potential data subsets of interest, the average access time for a particular statistical measure can be greatly reduced. Hence, a statistical program need only search the much smaller pointer files for information rather than searching the whole data base. Further, since within each pointer file the data is ordered by increasing PARM number, then only a minimum number of accesses of the main data file are necessary on a particular run.

Two things should be noted about this data base organization. First, the presorting of this data is a non-trivial computational task, involving many hours of computer sorting. This data base itself, therefore, is an important output of this effort, and may be used in the future for many classes of studies. Second, due to time constraints, DCEC was unable to make available enough information concerning the PARM data to take full advantage of this data base. Hence, the statistical resolving power afforded by this data base is better than that achieved by this study. Details of how the analysis could be improved is given later in this section.

2.4.2 The Statistical Analysis

The objective measures used in this study are shown in Table 2.11. The measures involved are essentially all the spectral distance measures used in the controlled distortion study (Section 2.3) plus

1.	D_1 LOG LPC
2.	D_1 LOG LPC GAIN WEIGHTED
3.	D_2 LOG LPC
4.	D_2 LOG LPC GAIN WEIGHTED
5.	D_4 LOG LPC
6.	D_4 LOG LPC GAIN
7.	D_2 LINEAR
8.	D_2 LINEAR GAIN WEIGHTED
9.	D_1 CEPSTRUM
10.	D_1 CEPSTRUM GAIN WEIGHTED
11.	D_2 PARCOR
12.	D_2 FEEDBACK
13.	D_2 AREA
14.	D_2 POLE LOCATION
15.	D_2 ENERGY RATIO

Table 2.11 Objective Measures Used in the PARM Correlation Study.

one additional measure which has had some attention in the literature [2.38].

The speech data used for this study was twelve sentences for each of two speakers (LL and CH) for each of the systems of Table 2.11. After the measures were applied, the statistical analysis performed was identical to that done for the controlled distortion tests.

In the correlation study, the categories recognized were "SUBJECT" and "SPEAKER." If the information had been available as to exactly which sentence was involved in which PARM, then "SENTENCE" could have been a category, increasing the degrees of freedom by approximately a factor of six. The correlation coefficients calculated were from

$$\rho = \frac{1}{K} \sum_{\text{subjects}} \sum_{\text{speakers}} \sum_{\text{systems}} \rho_a \qquad 2.25$$

where

$$\rho_a = \left(\frac{X_a - \bar{X}_s}{\hat{\sigma}_s}\right)\left(\frac{D_a - \hat{\bar{D}}}{\hat{\sigma}_D}\right) \qquad 2.26$$

where "a" is the condition including subject, speaker, and system, D_a is the distortion measure for that system, $\hat{\bar{D}}$ is the estimate of $\bar{D}$, X_a is the subjects response to condition "a", $\bar{X}_s$ is the average response for that subject over all systems, $\hat{\sigma}_s$ is the sample standard deviation for the subject "s," and $\hat{\sigma}_D$ is the sample standard deviation for the objective distortion measures.

In order to understand how these results are tabulated, it is first necessary to understand how results from the objective measures can be used to predict results from subjective tests.

The most straightforward way of deriving an estimate of the subjective quality is now given. Since both the subjective and objective measures for quality are means of a large number of independent estimates, then their marginal probability distribution functions are asymptotically normal, and, by the Bivariate Central Limit theorem, the joint probability distribution function is given by the Bivariate normal distribution:

$$f(X,D) = \frac{1}{2\pi\sigma_X\sigma_D\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{X-\bar{X}}{\sigma_X}\right)^2 - \frac{2\rho(X-\bar{X})(D-\bar{D})}{\sigma_X\sigma_D} + \left(\frac{D-\bar{D}}{\sigma_D}\right)^2\right\}\right], \tag{2.27}$$

where X is the subjective measure, D is the objective measure, σ_X is the variance of the subjective measure, σ_D is the variance of the objective measure, and ρ is the correlation coefficient. For this case, the minimum variance unbiased estimator of X from D is given by

$$X = \bar{X} + \frac{\rho\sigma_X}{\sigma_D}(D-\bar{D}) \tag{2.28}$$

where the variance of this measure is given by

$$E(X - E(X|D))^2 = \sigma_X^2(1 - \rho^2) \ . \tag{2.29}$$

If $\bar{X}$, $\bar{D}$, σ_X, σ_D, and ρ were known, this problem would be solved, since this is enough information to calculate confidence intervals on X or to do null hypothesis testing between systems. However, estimates for these quantities, called $\hat{\bar{X}}$, $\hat{\bar{D}}$, $\hat{\sigma}_X$, $\hat{\sigma}_D$, and $\hat{\rho}$, must be used instead, and these quantities are random variables themselves. Hence, the p.d.f. (Probability Distribution Function) is no longer normal, and is, in

general, very difficult to calculate in closed form.

However, considering the problem from the point of view of regression analysis theory offers additional information. The form of the linear regression estimation is given by

$$X = \beta_1 + \beta_2 D \ . \qquad 2.30$$

From the Gauss-Markov Theorem [2.40], the least squares estimate is the unbiased minimum variance estimate for X, and for this case (this is really an LPC analysis)

$$\hat{\beta}_2 = \frac{\sum_{j=1}^{N} X_j D_j - \left(\sum_{j=1}^{N} X_j\right)\left(\sum_{j=1}^{N} D_j\right)}{\sum_{j=1}^{N} D_j^2 - \left(\sum_{j=1}^{N} D_j\right)^2} = \frac{\hat{\rho}\hat{\sigma}_X}{\hat{\sigma}_D} \qquad 2.31$$

and

$$\hat{\beta}_1 = \frac{1}{N}\left(\sum_{j=1}^{N} X_j - \beta_2 \sum_{j=1}^{N} D_j\right) = \hat{\bar{X}} - \frac{\hat{\rho}\hat{\sigma}_X \hat{\bar{D}}}{\hat{\sigma}_D} \ . \qquad 2.32$$

Two points should be made here. First, these results show that the minimum variance unbiased estimator of X from D is gotten by using the minimum variance unbiased estimations for $\bar{D}$, $\bar{X}$, σ_X, σ_D, and ρ in Equation 2.28. Second, it should be noted that under a mild set of conditions easily met by the tests here, that four conditions hold: (1) a minimum variance unbiased estimate for σ_X^2, the variance in our approximation of the subjective quality, is given by

$$\hat{\sigma}_X^2 = \frac{1}{N-2} \sum_{j=1}^{N} (X_i - \hat{\beta}_1 - \hat{\beta}_2 D_i)^2 \ ; \qquad 2.32$$

(2) minimum variance unbiased estimates for the variance in $\hat{\beta}_1$ is given by

$$\hat{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}^2_X \left(\frac{1}{N} + \frac{\bar{\hat{X}}^2}{\sum_{i=1}^{N} (X_i - \bar{\hat{X}})^2}\right) ; \qquad 2.33$$

(3) a minimum variance unbiased estimate for the estimate for $\hat{\beta}_2$ is given by

$$\hat{\beta}_2 = \frac{\hat{\sigma}^2_X}{\sum_{i=1}^{N} (X_i - \bar{\hat{X}})^2} ; \qquad 2.34$$

and (4) the estimates for β_1 and β_2 ($\hat{\beta}_1$ and $\hat{\beta}_2$) are normal distributed, formed from $\hat{\sigma}^2_X/\sigma^2_X$, $\hat{\sigma}^2_{\hat{\beta}_1}/\sigma^2_X$, and $\hat{\sigma}^2_{\hat{\beta}_2}/\sigma^2_X$ are χ^2 distributed, and all five estimates are independent. These four points give all of the statistical power necessary to do all the hypothesis testing and confidence interval estimation which is normally associated with statistical testing and estimation. For example, if a confidence interval for β_1 was desired, it is only necessary to note that $(\beta_1 - \hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1})$ is t distributed, and the confidence interval is given by

$$\hat{\beta}_1 - U_{\alpha(N-2)}\hat{\sigma}_{\hat{\beta}_1} < \beta_1 < \bar{\beta}_1 - L_{(N-2)}\hat{\sigma}_{\hat{\beta}_1} , \qquad 2.35$$

where $U_{\alpha N-2}$ and $L_{\alpha N-2}$ are the upper and lower significance limits for a t distributed ($\mu = 0$, $\sigma = 1$) for N-2 degrees of freedom and probability α.

There are really two questions which these tests seek to answer.

First, assume that the estimates we have for correlations, means, and variance are exactly correct, what would then be the confidence intervals on our estimates of X? This question seeks to ascertain the potential of the objective measures used here to predict subjective results. Second, considering all the distorting factors in our analysis, especially our errors, in estimating β_1 and β_2, what then is the resolving power of our test? These questions address the usable resolving power of subjective acceptability estimates based on the analysis performed so far. The answer to the first question can be addressed by applying equation 2.29 to the estimate of the correlations (equation 2.25) of the correlation coefficients. The answer to the second question can be observed by applying equation 2.32 to the data.

2.4.3 The Experimental Results

The correlation studies described above were carried out on three sets of the data: all the systems; only the vocoder systems (LPC and channel vocoders); and only the waveform coders. The results for the three studies are given in Tables 2.12, 2.13, and 2.14, respectively. Several points should be made here. First, the correlation coefficients for a number of measures are quite high, some as high as .83. The "BEST" measures seem to be gain weighted spectral distance measures, as expected. Second, however, note that the estimated standard deviations are somewhat larger than desirable. This indicates that more data should be used to better establish these results. Third, note that much better results are obtained for the small subclasses than for the whole. This indicates that these measures work best if the systems being tested are preclassified according to the type of distortion expected.

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.76	10.24	22.24
D_1 LOG LPC GAIN WEIGHTED	-.79	8.13	16.13
D_2 LOG LPC	-.78	8.85	16.71
D_2 LOG LPC GAIN WEIGHTED	-.81	7.21	13.3
D_4 LOG LPC	-.73	14.31	24.12
D_4 LOG LPC GAIN WEIGHTED	-.78	8.31	16.3
D_2 LINEAR LPC	-.61	17.21	30.9
D_2 LINEAR LPC	-.66	13.21	27.1
D_1 CEPSTRUM	-.79	7.64	14.91
D_1 CEPSTRUM GAIN WEIGHTED	-.81	6.98	13.91
D_2 PARCOR	-.55	22.1	40.7
D_2 FEEDBACK	-.23	37.1	61.2
D_2 AREA	-.76	12.41	21.6
D_2 POLE LOCATION	-.25	21.6	40.7
D_2 ENERGY RATIO	+.78	9.2	18.3

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assuming $\hat{\rho}=\rho$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 2.12 Results of Correlation Study For Total Set of Systems

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.79	8.13	14.23
D_1 LOG LPC GAIN WEIGHTED	-.81	7.15	12.2
D_2 LOG LPC	-.79	8.27	18.3
D_2 LOG LPC GAIN WEIGHTED	-.83	6.63	13.4
D_4 LOG LPC	-.77	8.95	18.1
D_4 LOG LPC GAIN WEIGHTED	-.81	7.29	14.9
D_2 LINEAR LPC	-.70	16.31	31.6
D_2 LINEAR LPC GAIN WEIGHTED	-.74	14.52	28.4
D_1 CEPSTRUM	-.81	7.52	13.72
D_1 CEPSTRUM GAIN WEIGHTED	-.83	6.81	13.14
D_2 PARCOR	-.61	18.22	34.31
D_2 FEEDBACK	-.33	29.2	43.21
D_2 AREA	-.78	10.21	21.21
D_2 POLE LOCATION	-.36	36.3	61.3
D_2 ENERGY RATIOS	+.80	7.82	14.9

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assume $\rho=\hat{\rho}$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 2.13 Results of Correlation Study Using Only Vocoders.

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.79	8.23	14.12
D_1 LOG LPC GAIN WEIGHED	-.80	7.91	13.98
D_2 LOG LPC	-.78	9.41	18.91
D_2 LOG LPC GAIN WEIGHTED	-.82	6.78	12.21
D_4 LOG LPC	-.76	12.2	24.31
D_4 LOG LPC GAIN WEIGHTED	-.80	7.98	18.32
D_2 LINEAR LPC	-.73	14.23	29.31
D_2 LINEAR LPC GAIN WEIGHTED	-.75	12.9	26.21
D_1 CEPSTRUM	-.79	9.21	18.51
D_1 CEPSTRUM GAIN WEIGHTED	-.81	6.91	12.91
D_2 PARCOR	-.58	27.4	42.95
D_2 FEEDBACK	-.21	40.2	51.2
D_2 AREA	-.74	18.4	40.91
D_2 POLE LOCATION	-.31	29.6	51.9
D_2 ENERGY RATIO	+.76	16.3	33.6

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assuming $\rho=\hat{\rho}$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 2.14 Results of Waveform Coder Using Only Waveform Coders

These are certainly encouraging results. With measures as highly correlated as these, there is good expectation of creating a viable objective quality test. However, the relatively large estimated standard deviations in the estimates which include all statistics indicate more data must be processed to increase the resolving power of these tests to a maximum.

2.5 Summary and Areas for Future Research

The major results of this study can be summarized as follows.

(1) A number of objective quality measures, particularly spectral distance metrics, offer considerable promise in predicting subjective quality results.

(2) Some of the measures tested are clearly better than the others. The best are the gain weighted D_2 log LPC spectral distance measure and the gain weighted cepstral measure. These two measures are highly correlated with each other [2.35].

(3) Several measures do consistently poorly. Two of these are the D_2 feedback coefficient measure and the D_2 pole location measure. The pole location measure would probably improve if some sort of formant extraction was attempted.

(4) The D_2 area measure did quite well. This is interesting since it is so computationally compact.

(5) Gain weighting gave a slight, but consistent, improvement in the subjective-objective correlations.

(6) Based on the values of $\hat{\rho}$ obtained in this study, the potential for using several of the measures for predicting subjective scores is good. However, it should be noted that, even if $\rho=\hat{\rho}$, the resolving power of these tests falls short (by approximately a power of 2-2.5) of the subjective tests themselves. However, subjective and

objective measures may be combined to improve resolution. This is easily done so long as the number of subjective tests used warrants the use of the Bivariate Normal Distribution.

(7) The resolving power of the actual tests which resulted from this study are nowhere near as good as the "potential" resolving power. This is because the resolving power of the tests in this study on $\hat{\rho}$ was not good enough. This could be improved by doing a lower level correlation between a subject's response and the objective measure for the exact sentence used, and by using a larger protion of the PARM data base as part of the study. It should be noted, however, that although it is interesting to speculate on the improvement in the estimates of $\hat{\rho}$ that further testing would accomplish, no results should be assumed until the testing is complete.

The results of this study offer a number of areas for future research. Some of these are listed below.

(1) An obvious extension to this study would be to extend the portion of the PARM data base used in this study. This might well improve its estimates for $\hat{\rho}$.

(2) Statistically improved results may also obviously be obtained by finding measures which are more highly correlated with subjective results. One approach is to simultaneously attempt to better understand the parametric factors involved in human quality acceptance, as has been attempted in the "QUART" and "DAM" tests, and to develop objective measures which are highly correlated with the important parametric subjective measures.

(3) Improvements are possible in the particular objective measures used in the correlation studies. For example, Makhoul [2.13] suggests several forms of frequency weighting in LPC spectral distance measures which might be used to improve subjective-objective correlation.

REFERENCES

[2.1] "The Philosophy of PCM," B. M. Oliver, J. R. Pierce, and C. E. Shannon, Proceedings of The IRE, 1948.

[2.2] "Adaptive Quantization with a One Work Memory," N. S. Jayant, Bell Sys. Tech. J., vol. 52, 1973.

[2.3] "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," N. S. Jayant (Proceedings of the IEEE, May 1974).

[2.4] "Design and Implementation of an Adaptive Delta Modulator," N. S. Jayant, P. Cummiskey and J. L. Flanagan (Proceedings IEEE Int. Conf. Speech Communication, Boston, MA, April 1972).

[2.5] "Adaptive Delta Modulator with a One-Bit Memory," N. S. Jayant, Bell Sys. Tech. J., vol. 49, March 1970.

[2.6] "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems," M. D. Paeb and T. H. Glisson (IEEE Transactions on Communications, April 1972).

[2.7] "Adaptive Differential PCM Speech Trarsmission," T. P. Barnwell, A. M. Bush, J. B. O'Neal, P. W. Stroh, RADC-TR-74-177, Final Report, July 1974.

[2.8] "Adaptive Predictive Coding of Speech Signals," B. S. Atal and M. R. Schroeder (Bell System Technical Journal, October 1970).

[2.9] "Adaptive Transform Coding of Speech," Peter Noll, Personal Communication.

[2.10] "Adaptive Predictive Speech Coding Based on Pitch-Controlled Interruption/Reiteration Techniques," A. H. Frei, H. R. Schindler, and E. von Felten (Conference Record, 1973 IEEE International Conference on Communications, June 1973).

[2.11] "Speech Analysis and Synthesis by Linear Prediction of the Speech Waveform," B. S. Atal and S. L. Hanauer, J. Acoust. Soc. Amer., vol. 50, 1971.

[2.12] "Analysis Synthesis Telephony Based on the Maximum Likelihood Method," F. Itakura and S. Saito, Proc. Sixth Int. Congr. Acoust., 1968.

[2.13] "Linear Prediction: A Tutorial Review," J. Makhoul, Proc. IEEE, vol. 63, 1975.

[2.14] "Remaking Speech," H. Dudley, J. Acoust. Soc. Am., vol. 11, 1939a.

[2.15] "Dynamic Encoding as Applied to a Channel Vocoder," IEEE Trans. Comm. Syst., vol. 11, 1963.

[2.16] "Phase Vocoder," J. L. Flanagan and R. M. Golden, Bell System Tech. J., vol. 45, 1966.

[2.17] "Application of Digital Signal Processing to the Design of a Phase Vocoder," R. W. Schafer and L. R. Rapiner, 1972 Conf. on Speech Communication and Processing, 1972.

[2.18] "A Speech Analyzer and Synthesizer," W. A. Munson and H. C. Montgomery, JASA, Vol. 22, 1950.

[2.19] "Computer Simulation of a Format-Vocoder Synthesizer," JASA, Vol. 35, 1962.

[2.20] "Speech Analysis-Synthesis System Based on Homomorphic Filtering," A. V. Oppenheim, JASA, Vol. 45, 1969.

[2.21] "Homomorphic Analysis of Speech," A. V. Oppenheim and R. W. Schafer, IEEE Trans. Audio and Electro Acoust., AU-16, 1968.

[2.22] Speech Analysis, Synthesis, and Perception, J. L. Flanagan, Springer-Verlag, 1972.

[2.23] "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report AFSC No. F19628-70-C-0182, 1973.

[2.24] "Methods of Predicting User Acceptance of Voice Communications Systems," W. D. Voiers, Final Report DCA100-74-C-0056, July 1976.

[2.25] Digital Signal Processing, A. V. Oppenheim and R. W. Schafer, Prentice-Hall, 1975.

[2.26] Theory and Applications of Digital Signal Processing, L. R. Rabiner and B. Gold, Prentice-Hall, 1975.

[2.27] Sound Patterns of English, N. Chomsky and M. Halle, Haper & Row, New York, 1968.

[2.28] "Speech Synthesis by Rule," J. N. Holmes, I. G. Mattingly, and J. N. Shearme, Language and Speech, vol. 7, 1964.

[2.29] "Speech Synthesis by Rule: A Acoustic Domain Approach," L. R. Rabiner, Bell System Tech. J., vol. 47, 1968.

[2.30] "On Vowel Durations in English," A. S. House, JASA, vol. 33, 1961.

[2.31] "Duration and Intensity ARS Physical Correlates of Linguistics Stress," JASA, vol. 27, 1955.

[2.32] "Acoustical Correlates of Stress," J. Morton and W. Jassem, vol. 8, 1965.

[2.33] "The Quefrency Analysis of Time Series For Echos," B. P. Bogert, M. Healy, and J. W. Tukey, Proc. Symp. Time Series Analysis, 1963.

[2.34] "Echo Removal by Generized Linear Filtering," R. W. Schafer, IEEE NEREM Record, 1967.

[2.35] "Distance Measures for Speech Processing," A. H. Gray, Jr., and J. D. Markel, IEEE Trans. ASSP, vol. 24, 1976.

[2.36] "Pitch and Voicing in Speech Digitization," T. P. Barnwell, J. E. Brown, A. M. Bush, and C. R. Patisaul, DCA Research Report No. E-21-620-74-BU-1, August 1974.

[2.37] "Nonrecursive Digital Filter Design Using the IO-SINH Window Function," J. F. Kaiser, Proc. 1974 IEEE Sym. Cir. and Sys., 1974.

[2.38] "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE ASSP, vol. 23, February 1975.

[2.39] "Towards Perceptually Consistent Measures of Spectral Distance," J. Makhoul and W. Russel, 1976 IEEE Int. Conf. on Acoust., Speech, and Signal Processing, 1976.

[2.40] Mathematical Statistics, Samule S. Wilks, John Wiley, 1962.

CHAPTER 3

SUBJECTIVE PREDICTION OF USER PREFERENCE

3.1 Introduction

A crucial issue in the design and implementation of a digital voice communication system is the prediction of user acceptability. Even if the many other system design criteria are resolved and a good engineering solution found, the system will foil unless people use it. People will use it only if they find it highly acceptable on the basis of their current telecommunications alternatives.

Speech testing has been categorized as quality testing or intelligibility testing. The term preference testing or acceptability testing really supercedes both terms, not as a replacement for either, but as a combination of the essential features of each. That is, preference is assumed to be based on a sufficient combination of quality and intelligibility to determine relative user acceptability. It must be recognized here that 100% intelligibility may be yet of unacceptable quality and hence of low preference, just as pleasant but unintelligible speech is of low preference.

Just as with quality and intelligibility testing, preference testing can be implemented with a wide variety of strategies or methodologies. The test may be subjective, objective, parametric, isometric, based on absolute or relative scales, with an infinite variety of organizations. Fortunately, much work has been done in the testing of speech, so that we do not need to begin from scratch.

In this chapter we will consider subjective testing. Objective

testing, another phase of this effort, is considered in Chapter 2.

3.2 Subjective Testing Philosophies

Subjective testing procedures are based on drawing from a population of potential system users, i.e. subjects their reaction to the speech produced by a digital speech transmission system. These reactions must be quantified somehow and are then averaged, or processed, according to established statistical principles to arrive at a measure of user acceptance or preference. The basic testing philosophies can be listed as follows:

Iso-Preference Testing - involves the use of a known, agreed upon reference signal condition for use as a comparison in judging an unknown. The agreed upon conditioning must be parameterized so that the unknown or test signal can be found equally acceptable to an adjustment of the parameter set. This procedure then yields the judgement that a given signal is as acceptable as some reference condition.

Relative Preference Testing - involves comparisons, done independently, with each of several reference conditions. The reference conditions are used to establish a scale of preference, and an unknown signal can then be ranked on this scale. The subjective scale of the references must be agreed upon a priori.

Absolute Preference Testing - methods require the subjects performing the test to give an absolute numerical evaluation to the properties described in the test format. Properties tested can be selected to describe various features of interest.

Isometric Testing for user preference calls for a direct evaluation of preference from the test subjects. Each subject makes his evaluation against the background of his total experience and personal biases, and including any local or instantaneous bias with fatigue or irritability effects built into his response.

Parametric Testing asks the test subject to make judgements with respect to specific features of the speech signal under consideration. The test format has then the flexibility of later weightings of feature judgements to achieve a measure of acceptability which is more independent of the individual subject's biases. The appropriate weightings must be agreed upon in the final resolution of test data however.

The most recent application of these philosophies has resulted in the PARM test and the QUART test [3.1] and more recently in the DAM test [3.2].

In the PARM test (Paired Acceptability Rating Method) an isometric approach is used. However, since systems being tested are presented to the subjects in a carefully chosen ordering, paired comparisons can be abstracted from the test results or on a posteriori basis. To reduce the effects of extremes of responses typical in isometric testing, the listerns are asked to judge two reference or anchor conditions, one "good" and one "bad" anchor. Anchor responses are then used to normalize other responses within and across listeners. Details of the testing organization and exhaustive analysis of results are found in [3.1].

In the QUART test (Quality Acceptance Rating Test) the parametric philosophy is followed, with an isometric measure of overall acceptability

included as well. The listener is asked to score each system under test with respect to a family of features and to give his overall reaction. Extensive analysis of this approach is also well documented [3.1].

An outgrowth of the background of subjective testing of speech in general and of experience with PARM and QUART in particular, after substantial further requirement in the choice of a family of features to use in direct response solicitation, is the DAM test (Diagnostic Acceptability Measure).

The DAM test acquires ratings on perceptual features which have been selected after extensive experience with QUART as those features closely correlated with overall acceptability, nearly orthogonal to each other, and directly related to specific system functions or to system operating environment conditions. In addition, the feature set thus extracted is small enough to allow efficient and reasonable subjective testing to be accomplished. The DAM test is still evolving, but is nearing a final form. Although it is not yet documented in the literature, the test has been the subject of substantial interaction between the speech research group at Georgia Tech and the group at Dynastat. These discussions have been conducted in visits by A. M. Bush and T. P. Barnwell to Dynastat and by W. D. Voiers to Georgia Tech. A detailed description of the DAM test is included as Appendix A of this report.

3.3 Statistical Testing Procedures

In subjective testing, as mentioned earlier, an essential aspect of the test implementation is the statistical processing of the data, i.e. responses from listeners or subjects, to obtain an average rating of the system or system feature under test. Even though the field of

statistics is well documented, both in the scientific literature and in textbook and reference book formats, it is our feeling that some exposition here may be worthwhile. Our point of view (necessarily!) is that of the communications engineer with a background in probability, random variables, and stochastic processes, who feels he should therefore know all about statistics until he reads a little in the area.

In order to apply statistics to the results of subjective testing, one must either base the statistics on assumptions regarding the underlying distributions of the individual listener responses, the parametric approach, or assume that these underlying distributions are unknown and work within, for example, ranking statistics, the nonparametric approach. The parametric approach is treated from a theoretical approach in many places: our favorites are Wilks [3.3], and Cramer [3.4]. The nonparametric approach is also extensively treated, but our favorite here is Hajek [3.5]. For applications with a minimum of theory, a good reference among a great many possible choices is Winer [3.6] or Siegel [3.7] for parametric or nonparametric tests, respectively.

In the parametric approach, the most common assumption regarding the distribution of the listener responses is that they are all Gaussian. Hypotheses with respect to common means and/or variances under test conditions can then be set up and inferences drawn by comparisons with standardized tables.

3.3.1 Distributions

The key distributions are summarized below for convenience.

Chi Square

Let X_i, i=1,...,n be independent, identically distributed Gaussian random variables, each with zero mean and unit variance. Then

$$\chi^2 = \sum_{i=1}^{n} x_i^2 \tag{3.3.1}$$

is a new random variable, with a distribution called Chi-square with n degrees of freedom. The probability density function is given by

$$f_{\chi^2}(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x \geq 0 \\ & x < 0 \end{cases} \tag{3.3.2}$$

F-Distribution

Let $X_1,\ldots,X_n$, and $Y_1,\ldots,Y_m$ be n+m independent, identically distributed Gaussian random variables each with zero mean and unit variance. Then the ratio

$$F = \frac{\frac{1}{m}\sum_{i=1}^{m} Y_i^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \tag{3.3.3}$$

is a random variable with a distribution called the F-distribution, with parameters m and n. The probability density function is

$$f_F(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} (\frac{m}{n})^{\frac{m}{2}} x^{\frac{m}{2}-1} (1 + \frac{mx}{n})^{-\frac{m+n}{2}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Student's Distribution

Let $X_0, X_1, \ldots, X_n$ be independent identically distributed Gaussian random variables each with zero mean and unit variance. Let

$$t = \frac{x_0}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}} \tag{3.3.5}$$

Then t is a random variable which has a distribution called the Student's distribution with parameter n. The probability density function is

$$f_t(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}} \tag{3.3.6}$$

Studentized Range Statistic

Let $X_1, \ldots, X_k$ be independent identically distributed Gaussian random variables each with zero mean and unit variance. Define a random variable Z as

$$Z = \max_i(x_i) - \min_i(x_i) \tag{3.3.7}$$

as shown in Figure 3.1. The probability density function of Z is

$$f_Z(x) = \begin{cases} k(k-1)\int_{-\infty}^{+\infty} (F_X(\xi)-F_X(\xi-x))^{k-2} f_X(\xi) f_X(\xi-x)\,dx & x > 0 \\ 0 & x < 0 \end{cases} \tag{3.3.8}$$

where $F_X(\cdot)$ is the Gaussian cumulative distribution function and $f_X(\cdot)$ is the Gaussian probability density function, both for zero mean, unit variance Gaussian random variables. This function is not available in closed form unless k=2. Some points of the cumulative distribution function for Z have been tabulated. See for example the tables of Winer [3.6]. For a derivation of (3.3.8), see Appendix B of this report.

3.3.2 Estimation

We consider now some commonly used estimates of statistical parameters.

Mean

Let $X_1,\ldots,X_n$ be independent identically distributed random variables each with mean μ and variance σ^2. Then

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.3.9}$$

is called the sample mean. It is an unbiased estimate of the mean of

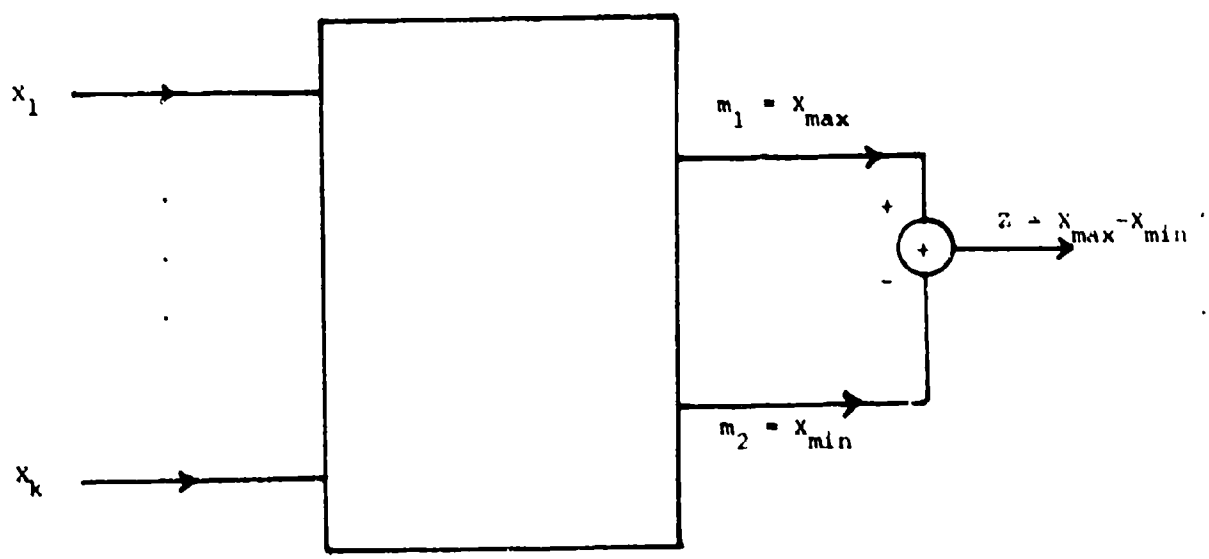


FIGURE 3.1

Generation of Studentized Range Statistic.

the x_i's:

$$E[\bar{x}] = \mu \tag{3.3.10}$$

$$Var[\bar{x}] = \frac{\sigma^2}{n} \tag{3.3.11}$$

Variance

For $X_1,\ldots,X_n$ independent identically distributed random variables, each with mean μ and variance σ^2, the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.3.12}$$

is an unbiased estimate of the variance of the x_i's, with

$$E[s^2] = \sigma^2 \tag{3.3.13}$$

$$Var[s^2] = \frac{1}{n}(u_4 - \frac{n-3}{n-1}\sigma^4) \tag{3.3.14}$$

where u_4 denotes the fourth central moment. If the x_i's are Gaussian as well, then $\bar{x}$ and s^2 are best mean square estimates and are independent random variables. Also, in this case,

$$t = \frac{\sqrt{n}\ (\bar{x}-\mu)}{\sqrt{s^2}} = \sqrt{n}\ \frac{\bar{x}-\mu}{s} \tag{3.3.15}$$

is a random variable with the student's distribution with (n-1) degrees of freedom.

3.3.3 Analysis of PARM Data

As an example of the application of the above results, let us consider the problem of analysis of the PARM data. Let

R_{ijk} = the response of listener i to system j on the k^{th} presentation

For a particular PARM module of data, we have

$1 \le i \le L$ = the number of listeners in the module

$1 \le j \le M$ = the number of systems in the module

$1 \le k \le 10S = T$ = the number of times a system is presented in a module, where S is the number of speakers in the module.

For example, L=10, M=6 including anchors, S=3, T=30 might be a set of parameters, with 1800 total responses in the module.

Let

$$\overline{R_{ij}} = \frac{1}{T} \sum_{k=1}^{T} R_{ijk} \qquad (3.3.16)$$

$$\overline{R_{i}} = \frac{1}{M} \sum_{j=1}^{M} \overline{R_{ij}} \qquad (3.3.17)$$

$$\overline{R_{j}} = \frac{1}{L} \sum_{i=1}^{L} \overline{R_{ij}} \qquad (3.3.18)$$

$$\bar{R} = \frac{1}{MLT} \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{T} R_{ijk} \qquad (3.3.19)$$

$$\hat{\sigma}^2 = \hat{\sigma}^2_{total} = \frac{1}{MLT-1} \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{T} (R_{ijk} - \bar{R})^2 \qquad (3.3.20)$$

$$\hat{\sigma}^2_{sys} = \frac{1}{M-1} (LT \sum_{j=1}^{M} (\bar{R}_j - \bar{\bar{R}})^2) \qquad (3.3.21)$$

$$\hat{\sigma}^2_{error} = \frac{1}{M(LT-1)} (\sum_{j=1}^{M} (\sum_{i=1}^{L} \sum_{k=1}^{T} (R_{ijk} - \bar{R}_j)^2)) \qquad (3.3.24)$$

Then, combining results, we have

$$\hat{\sigma}^2_{total} = \frac{M-1}{MLT-1} \hat{\sigma}^2_{sys} + \frac{M(LT-1)}{MLT-1} \hat{\sigma}^2_{error} \qquad (3.3.25)$$

Now, if $\hat{\sigma}^2_{sys} = \hat{\sigma}^2_{error}$, that is, if the different systems themselves contribute no systematic differences to the variance, then

$$\hat{\sigma}^2_{total} = \hat{\sigma}^2_{sys} + \hat{\sigma}^2_{error} \qquad (3.3.26)$$

The F-test is used to test the hypothesis that $\hat{\sigma}^2_{sys} = \hat{\sigma}^2_{error}$, by forming the ratio of these variables, assuming that the Gaussian assumptions hold, and utilizing tabulations of the cumulative distribution of the F variable under the hypothesis. If the ratio is outside predetermined bounds, the test is said to hold, that is, the two variances are not the same. Otherwise, there is no conclusion. From the point of view of statistical hypothesis testing, we test the hypothesis {systems contribute no systematic difference}. If F is too large, we reject the

hypothesis. This amounts to considering the hypothesis against a specified false alarm probability, and not giving any other measure of performance.

For a comparison between pairs of means, one can use the studentized range statistic as

$$q_{\alpha,R,f} = \frac{\bar{R}_{(j)} - \bar{R}_{(j')}}{\sqrt{\frac{\sigma^2_{error}}{LT}}} \qquad (3.3.25)$$

where α is the desired quantile point of the cumulative distribution of the statistic, $R=(j)-(j')+1$, $2 \le R \le M$, is the number of steps between the $\bar{R}_j$'s being compared when all the $\bar{R}_j$'s are rank ordered, and $f=M(LT-1)$ - degrees of freedom of σ^2_{error}. When this test is organized in matrix form to facilitate the comparison of all means for significance of differences between pairs of means to level α of false alarm, the procedure is called the Newman-Keuls test. (See Winer [3,6] pp. 80-81).

3.3.4 Nonparametric Tests

In nonparametric testing, one declines to assume that the underlying statistics are Gaussian. Then one ranks the responses corresponding to their relative magnitude either signed or unsigned. If the conditions hypothesized give no systematic differences in responses, the rankings will be purely random, resulting in statistics which for two conditions may be derived fairly easily. Common two dimensional nonparametric tests resulting from various ranking procedures are the

Wilcoxon test, the Median test, the Van der Warden test, and the Kolmogorov-Smirnov test. Hajeck [3.5] describes each of these tests and gives underlying statistics for which each test is most powerful. Unfortunately, no uniformly most powerful test exists. In situations where underlying distributions may reasonably be assumed to be Gaussian, a parametric test will in general be best.

Nonparametric tests comparing more than two conditions are more difficult to compose than the comparisons of pairs of conditions as all the rank statistics are in the higher order case derived from multinomial as opposed to binomial type distributions. Although some references are made to such proceudres in Lehman [3.8], e.g. the Kruskal-Wallis test, no convenient generally accepted multidimensional nonparametric tests were found.

3.4 Conclusions and Recommendations

The following conclusions regarding subjective prediction of user preference are drawn primarily on the basis of data available from the analysis of the results of the PARM and QUART tests [3.1], from the discussions at Georgia Tech and at Dynastat with W. D. Voiers, and from the initial results of the DAM test.

3.4.1 Isometric Tests

In isometric tests such as PARM, the absolute rankings of system conditions by individual listeners will have a high variance due to individual listener idiosyncrasies and intralisteners variability, in addition to interlisterner variability. Although these effects can be

balanced out by extremely careful post-test processing of responses to establish common origins and scales within and across listeners. Such processing is, inevitably, subject to some criticism, as any smoothing of the data will also introduce some distortion of one kind as it reduces other effects. Smoothing, centering, and scaling was accomplished in the PARM tests based on the ratings and relative ratings of the anchors. Although more efficient anchoring and normalization procedures can clearly be devised, such tests will always suffer from high variability and hence require large groups of listeners and many trials and will always be subject to criticism due to post test normalization procedures.

3.4.2 Tests of Features

In order to devise an effective, efficient and reliable subjective test, it is necessary to narrow the scope of the question asked the system. That is, a more specific response than "Do you like this?" must be solicited. If the features of the speech which are perceptually most important in determining the overall user acceptability can be identified and quantified, than one can construct an acceptability rating with less variability within and across listeners.

This then becomes a problem of feature extraction. Two fronts or approaches to this problem can be found: (a) List all the conceivable descriptions of features. Test. Analyze the data with correlation analysis and try to find the features which are important empirically. (b) Based on extensive experience with various systems, select the most typical types of noises and degradations. Try to solicit responses along these particular features. Include effects of the environment such

as background noises. Feature selection using method (a) was used in QUART. Subsequent refinement using the ideas of (b) as well have led to the parameter sets of DAM. It is our judgment, based on the results of DAM, that the best available subjective preference testing procedure available now is DAM. It should be pointed out that until the extensive, expensive, detailed test results of PARM and QUART it was not possible to draw this conclusion; however, the detailed agreement of PARM and QUART, and the subsequent development of DAM leave no other conclusion.

3.4.3 Implementation of Subjective Tests

The monumental and time consuming tasks of conducting a subjective listening test can effectively be implemented for improved speed and efficiency by using an interactive computer to control the test, collect the data, and subsequently to analyze the test data.

3.4.4 Size of the Test

The numbers of listeners which must be used in a subjective testing procedure can be determined only after sufficient data is accumulated with a particular test methodology or algorithm to permit good estimation of the error variances. Then the number of responses must be selected to give an adequate resolution of the data to separate systems under test. Note that the required resolution also will depend on how different the systems to be resolved are on the scale of interest.

3.4.5 Speaker Selection

The number of speakers has been found in QUART and PARM to be less significant than previously thought, from the point of view of

statistical resolving power. However, from the point of view of system design, it is clear that some systems will be highly biased toward low pitched speech or moderately pitched speech, and perform quite poorly on high pitched speech or vice-versa. Hence, it is considered essential to use at least two, preferably three, speakers chosen to cover the expected range of pitches. This strategy will at least isolate quickly systems which will not, for example, respond to a female voice.

3.4.6 Overall Recommendations for Subjective Tests

The overall recommendation to come from this examination of subjective tests and test facilities is the development of an interactive computer based hardware facility for conducting a refined version of the DAM test.

REFERENCES

3.1 Voiers, W. D., "Methods of Predicting User Acceptance of Voice Communication Systems," Final Report, DCA Contract No. DCA-100-74-C-0056, July 15, 1976.

3.2 Voiers, W. D., "The Diagnostic Acceptability Measure Test," See Appendix A of this report.

3.3 Wilks, Sammuel S., Mathematical Statistics, John Wiley & Sons, 1962.

3.4 Cramer, H., Mathematical Methods of Statistics, Princeton University Press, 1946.

3.5 Hajek, Jaroslav, A Course in Nonparametric Statistics, Holden Day, 1969.

3.6 Winer, B. J., Statistical Principles in Experimental Design, McGraw-Hill, 1962.

3.7 Siegel, S., Nonparametric Statistics for the Behaviorial Sciences, McGraw-Hill, 1956.

3.8 Lehman, E. L., Non Parametrics: Statistical Methods Based on Ranks, Holden Day, 1975.

SELECTED BIBLIOGRAPHY IN SPEECH QUALITY TESTING

1. "IEEE Recommended Practice for Speech Quality Measurements," IEEE Trans. on Audio and Electroacoustics, Vol. AU-17, No. 3, Sept. 1969, pp. 225-246.

2. J. D. Carrol, J. J. Chang, "A New Method for Dealing with Individual Differences in Multidimensional Scaling," Bell Telephone Laboratories, Technical Memorandum, 69-1221-3, Jan 3, 1969.

3. J. D. Carroll, "Individual Differences and Multidimensional Scaling," Preprint, Bell Telephone Laboratories.

4. L. H. Nakatani, "Confusion-Choice Model for Multidimensional Psychophysics," Bell Telephone Laboratories, Technical Memorandum, 70-1234-2, March 30, 1970.

5. J. B. Kruskal, "Nonmetric Multidimensional Scaling," Bell Telephone Laboratories, Technical Publications, August, 1964.

6. M. R. Schroeder, "Reference Signal for Signal Quality Studies," Bell Telephone Laboratories, Technical Memorandum, 68-123-3, July 18, 1968.

7. L. H. Nakatani, B. J. McDermott, "Effect of Pitch and Formant Manipulations on Speech Quality," Bell Telephone Laboratories, Technical Memorandum, 72-1228-8; 72-1229-5, September 1, 1972

8. L. H. Nakatani, "Measuring the Ease of Comprehending Speech," Bell Telephone Laboratories, Technical Memorandum, 71-1234-5, June 2, 1971.

9. P. D. Bricker, "Application of INDSCAL to the Dimensionalization of Speech Sound Perception," Bell Telephone Laboratories, Technical Memorandum, 70-1221-24, August 25, 1970.

10. J. D. Carroll, M. Wish, "Multidemsional Scaling of Individual Differences in Perception and Judgment," Preprint, Bell Telephone Laboratories.

11. L. H. Nakatani, "Evaluation of Speech Quality Using Immediate Recall of Digit Strings," Bell Telephone Laboratories, Technical Memorandum, 70-1234-1, January 29, 1970.

12. Michael H. L. Hecker, C. E. Williams, "Choice of Reference Conditions for Speech Preference Tests," The Journal of the Acoustical Society of America, Vol. 39, No. 5, Part 1, 1966, pp. 946-952.

13. E. H. Rothauser, G. E. Urbanek, W. P. Pachl, "A Comparison of Preference Measurement Methods," The Journal of the Acoustical Society of America, Vol. 49, No. 4, Part 2, 1971, pp. 1297-1308.

14. B. E. Carpenter, S. H. Lavington, "The Influence of Human Factors on the Performance of a Real-Time Speech Recognition System," The Journal of the Acoustical Society of America, Vol. 53, No. 1, 1973, pp. 42-45.

15. Michael H. L. Hecker, Newman Guttman, "Survey of Methods for Measuring Speech Quality," Journal of the Audio Engineering Society, Vol. 15, No. 4, Oct. 1967, pp. 400-403.

16. W. D. Voiers, "Perceptual Bases of Speaker Identify," The Journal of the Acoustical Society of America, Vol. 36, No. 6, June 1964, pp. 1065-1073.

17. W. A. Munson, J. E. Karlin, "Isopreference Method for Evaluating Speech-Transmission Circuits," The Journal of the Acoustical Society of America, Vol. 34, No. 6, June 1962, pp. 762-774.

18. P. T. Brady, "A Statistical Basis for Objective Measurement of Speech Levels," The Bell System Technical Journal, Sept. 1965, pp. 1453-1486.

19. P. T. Brady, "Objective Measures of Peak Clipping and Threshold Crossings in Continuous Speech," The Bell System Technical Journal, April 1972, pp. 933-945.

20. W. D. Voiers, "The Present State of Digital Vocoding Technique: A Diagnostic Evaluation," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 2, June 1968, pp. 275-279.

21. J. Swaffield, D. L. Richards, "Rating of Speech Links and Performance of Telephone Networks," The Proceedings of the IEEE, Vol. 106, Part B., No. 26, March 1959, pp. 65-76.

22. D. L. Richards, J. Swaffield, "Assessment of Speech Communication Links," IEEE, Paper No. 2605 R, April 1958, pp. 77-89.

23. L. H. Nakatani, K. D. Dukes, "A Sensitive Test of Speech Communication Quality," The Journal of Acoustical Society of America, Vol. 53, No. 4, 1973, pp. 1083-1092.

24. B. J. McDermott, "Multidimensional Analyses of Circuit Quality Judgements," The Journal of the Acoustical Society of America, Vol. 45, No. 3, 1969, 774-781.

25. E. H. Rothauser, "Isopreference Method for Speech Evaluation," The Journal of the Acoustical Society of America, Vol. 44, No. 2, 1968, pp. 408-418.

26. J. R. Duffy, T. G. Giolas, "Sentence Intelligibility as a Function of Key Work Selection, University of Connecticut, Storrs, Connecticut, Feb. 1974.

27. V. E. McGee, "Semantic Components of the Quality of Processed Speech," Journal of Speech and Hearing Research, No. 6, pp. 310-323.

28. V. E. McGee, "Determining Perceptual Spaces for the Quality of Filtered Speech," The Journal of Speech and Hearing Research, No. 8, pp. 23-38., 1965.

29. D. W. Bell, E. James Kreul, "Reliability of the Modified Rhyme Test for Hearing," The Journal of Speech and Hearing Research, No. 15, 1972, pp. 287-295.

30. W. D. Voiers, C. P. Smith, "Diagnostic Evaluation of Intelligibility in Present-Day Digital Vocoders," Conference Record, 1972 Conference on Speech Communication and Processing, April, 1972, pp. 170-174.

31. C. B. Grether, R. W. Stroh, "Subjective Evaluation of Differential Pulse Code Modulation using the Speech 'Goodness' Rating Scale," Conference Record, 1972 Conference on Speech Communication and Processing, April, 1972, pp. 175-178.

32. D. J. Goodman, B. J. McDermott, L. H. Nakatani, "Subjective Evaluation of PCM Coded Speech," Conference Record, 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing, April, 1976, pp. 266-270.

4. A SUBJECTIVE COMMUNICABILITY TEST

4.1 Introduction

When judging the performance of highly intelligible speech communications systems, one approach is to apply an isometric subjective user acceptability test, such as the PARM. The hypothesis in such tests is that subjects can judge, from listening to speech segments played through the systems being tested, the overall expected acceptability of a system. The problem with these tests is that the subjects' responses represent a noisy measure of the actual acceptability of a system. In this context, the "ACCEPTABILITY" of a system is defined as the level to which complex communication tasks can be accomplished while using the system.

A model which states the problem more clearly is one which postulates a fixed cognitive resource available to a user of a communication system. As was discussed in Chapter 2, due to the multiplicity of acoustic cues for segmental and supersegmental features in speech, and due to a listener's immense knowledge of the phonemics, syntactics, and semantics of his language, a listener may well be able to understand speech which is very distorted. The problem is that to do so, he must utilize a large portion of his cognitive resource to just understanding what is being said. For a low quality system, therefore, this leaves him relatively less cognitive resource to apply to the communication task, making the communication task more difficult.

The definition of a "COMMUNICABILITY TEST," as used in this chapter, is any test which trys to measure a user performance on a communication task while using a communication system. The idea is to design tests in which users are not asked to rate systems, but rather are asked to perform some task in which the subjects' performance may be measured objectively. In order to be an acceptable communicability test, therefore, the test must meet several requirements. First, the communication task must be difficult enough so that a subject is using most of his cognitive resource in performing the task even with no system distortion. Second, a subject's performance on the task must be easy to measure. Third, the test must be inexpensive to administer because it has enough inherent resolving power to differentiate among the communications systems without excessive subject costs. Last, the test should not require the actual use of a communication system in the test, so that simulated systems may also be tested.

This chapter describes the design and testing of one such communicability test. Section 4.2 describes the design of the automated subjective data acquisition system used to administer the test. Section 4.3 describes the details of the test itself. Section 4.4 describes the data analysis done in the test. Section 4.5 describes the test results.

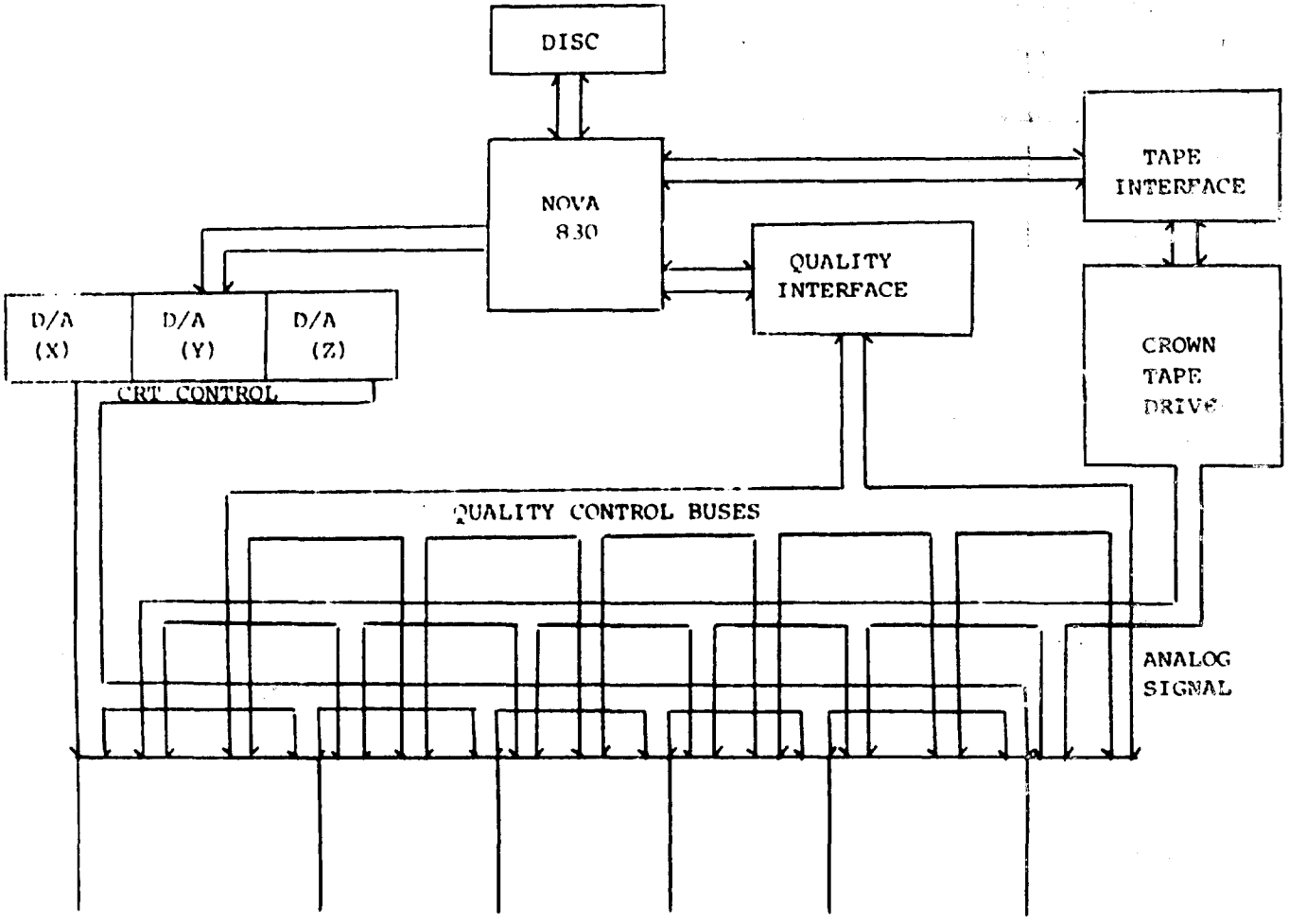
4.2 An Automated Speech Subjective Quality Testing Facility

One of the greatest sources of expense in performing subjective speech quality tests is the large amount of manual data handling required to prepare the test results for computer analysis. In order to reduce this source of expense, an automated subjective data acquisition system was developed at Georgia Tech.

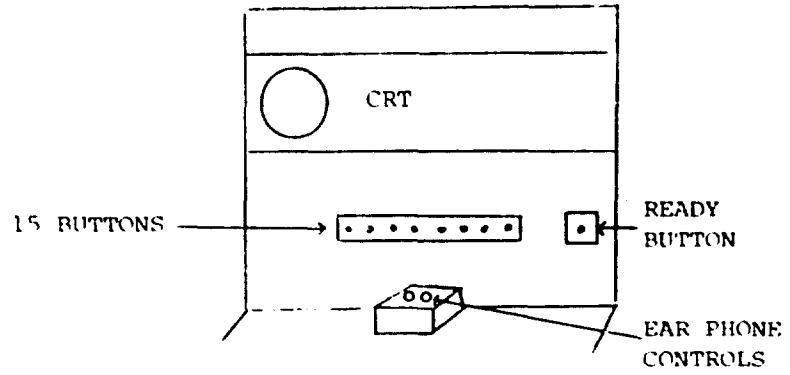
A diagram of the hardware portion of the subjective data acquisition system is shown in Figure 4.1. The system consists of six "STATIONS," each of which has an earphone control console, a CRT, and a total of 16 buttons, fifteen "DATA" buttons and one "CONTROL" button. The CRT is used for transmitting alphanumeric data to the subjects through the computer's D/A interfaces, while the buttons are used for collecting subject responses. The audio for the system is supplied by a Crown 800 analog tape recorder which is digitally controlled. In general, 1 kHz tones are placed one track of the analog tape to mark the ends of test sequences. These tones can be detected by the computer through a phase lock loop detector, and are used to accurately position the recorder.

In order to administer the test and collect the data, a multitask interpretive test control program, called "QUALGOL," was written. The QUALGOL language is summarized in Table 4.1, and has all the necessary elements (constants, variables, labels, loop control, arithmetics, etc.) for a simple computer language. Using the QUALGOL language, an experimenter can easily "PROGRAM," a large class of subjective tests on the quality testing facility. A program used for administering some of the tests performed during this study is given in Figure 4.2.

HARDWARE FOR QUALITY TESTING



SIX STATION QUALITY FACILITY



QUALITY STATION

Figure 4.1

TABLE 4.1
QUALGOL LANGUAGE

CONVENTIONS: V = VARIABLE
N = CONSTANT

VARIABLES: A-Z

COMMANDS:

C	CROWN	
	C(V)	RECEIVE FROM CROWN 1 = TONE 0 = NO TONE
	C(N)	SEND TO CROWN 1 FAST FORWARD 2 STOP 3 PLAY 4 RECORD 5 REWIND 0,6,7 NO-OP
D	DELAY	
	D(N)	DELAY N(.1 SEC) UNITS
DI	DISPLAY	
	D(N)	DISPLAY MESSAGE N
E	END	
G	GET RESPONSES	
	G(V)	GET V RESPONSES DECREMENT V TO ZERO
I	INCREMENT	
	I(V)	INCREMENT V BY ONE
J	JUMP	
	J(V,LABEL)	JUMP TO LABEL IF V=0
	J(@,LABEL)	JUMP TO LABEL
M	MESSAGE	
	(M(N,"...")	DEFINE MESSAGE
P	PRINT	
	P(V)	PRINT V
S	SET	
	S(V,N)	SET V TO N
T	TRACE	
	T	TRACE SWITCH
W	WAIT	
	W(N)	WAIT N UNITS

```
      (1,LISTEN@    TO@SAMPLE)
      M(2, )
      M(3,  MAKE@CHOICE    NOW)
      M(4,PLEASE@MAKE CHOICE@  NOW)
      M(5,NOW STUPID)
      S(E,-1ØØ)
      C(3)W(2)C(Ø)
LT    C(B)J(B,L1)J(@.L2)
L2    C(2)W(2)C(Ø)

LM    I(E)J(E,EN)DI(1)
      C(3)W(3)C(Ø)
L3    C(B)J(B,L4)J(@.L3)
L4    C(B)J(B,L4)J(@.L5)
L5    C(2)W(2)C(Ø)
      DI(2)W(1Ø)
      DI(3)S(C,1)G(C)W(3Ø)
      J(C,LM)DI(4)S(D,-1Ø)
L7    W(1Ø)J(C,LM)I(D)(D,L8)J(@,L7)
L8    S(D,-1Ø)DI(S)
L9    J(C,LM)W(1Ø)I(D)J(D,LM)J@,L9)
EN    END
```

FIGURE 4.2 AN EXAMPLE "QUALCOL" PROGRAM USED TO ADMINISTER THE COMMUNICABILITY TESTS

4.3 The Experimental Format

The communicability test format chosen for this study was a "Multiple Digit Recall" test similar to that studied by Naghtani at Bell Labs. In this format, sequences of random digits are first recorded by trained speakers, and then these utterances are played through various distorting systems. The resulting sequences are then played to subjects whose task is to "RECALL" the digits after a short (~ 1 sec.) wait. This test format meets all the basic criteria set forth in the introduction, since the task does not require a quality judgment on the part of the subjects, the test is simple to administer, and the test does not require the communication system being tested to be present.

The purpose of the study reported here was to study the usefulness of this test format for evaluating communication systems both from a resolution and cost point of view. It should be noted that this study was a relatively small portion of the total effort, and the results obtained should be considered preliminary in nature. The tests were performed as follows. First, strings of random digits were generated by the computer by a program which rejected all strings which had double digits, had more than two digits in ascending or descending sequence, or had more than two digits in ascending or descending alterate (2-4-6, etc.) sequence. Forty random sequences were generated in 6,7,8,9, and 10 digit lengths. Second, the digit strings were read into a high quality tape recording system by a trained announcer from the student broadcast radio station. The digits were read "as if there were a list," so that no internal groupings were imposed on the numbers. Third, the number strings were low pass

filtered to 3.2 kHz and digitized at 8 kHz to 12 bits resolution. The results were stored on three 2400 ft. 800 BPI, 9 track digital tapes.

In all, four sets of tests were performed. In the first "preliminary" test, undistorted data was played to subjects to try to determine an appropriate number of digits for the final tests. In all, the subjects listened to 200 sequences consisting of 40 each of 6,7,8,9, and 10 digit strings. As a result of this test, digit sequence lengths of 7 and 8 were chosen.

In the remaining three tests, distortions were applied to the number strings, and these were played to subjects. Each of these three tests tested the undistorted strings against three levels of easily perceivable distortions. In the first test, the distortions were white Gaussian noise at a SNR of 10 db, 8 db and 5 db. In the second test, the distortions were low pass filtering at 2.4 kHz cut-off frequency, 1.8 kHz cutoff frequency, and 1.2 kHz cutoff frequency. In the third test, the distortions were ADPCM waveform coder distortions at 24 kBPs, 16 kBPs, and 8 kBPs. Each set of distortions was played to 18 subjects for a total of 18x3x2x50=5400 responses.

4.4 The Data Analysis

The data analysis was done in three stages. First, the data is entered into a general data base. Second, a program called "VERIFY' examines the numbers for cases where the number of errors is greater than three, or where the errors meet a set of special conditions (reversals, dropped numbers, etc.). In each case, the experimenter can choose to omit the subject data. Third, a program called "SCORE" allows the analysis of the data base for the means and variances

necessary to use standard Student's-t analysis and analysis of variance techniques, and allows the calculation of extensive correlation sets.

In all, three types of scoring procedures were applied to the data. In the first procedure, each response string was scored to be either correct or not correct, and no note was made of the number of errors in the string. The score statistic for this method was the percentage of incorrect strings for each subject, for each distortion, and for each test.

In the second scoring procedure, each response string was matched to the correct string, and the score was taken to be the total number of incorrect digits. In this scoring procedure, all response strings with missing digits or response strings with the wrong number of digits were given a score of 4.

The third type of scoring was derived by classifying the types of digit errors in the response strings. It was found that the predominant type of error in the test was a two digit error obtained from interchanging two digits. In the third scoring procedure, such an inversion would be considered to be one error rather than two. Rules were compiled to handle inversions of more than two numbers as such cases appeared in the data.

For the following discussion, each scored result will be referred to by the designation X_{tsdm}, where t is test number (t = 1 for the additive noise test, t = 2 for the low pass filter test, and t = 3 for the ADPCM coding test), s is the subject number (18 per test,

$1 \leq s \leq S$, where $S = 18$), d is the distortion level (four for each test - three distortions and "clear" $1 \leq d \leq D$, where $D = 4$), and n is the number of results per subject ($1 \leq n \leq N$, where $N = 1$ for the first scoring, and $N = 10$ for the last two). For each test, analysis of variance was used to determine the significance of the entire test, while the Student's t statistic was used to determine statistical significance between distortions. In each test, the first 10 responses were considered to be "training" responses, and were not included in the results. The analysis of variance was performed by calculating the F statstic given by

$$F_t = \frac{\frac{1}{D-1} \sum_{d} (\bar{X}_{td} - \bar{X}_t)^2}{\frac{1}{D(SN-1)} \sum_{d} \sum_{s} \sum_{m} (X_{tsdm} - \bar{X}_{td})^2} \qquad (4.4.1)$$

and testing for significance using the appropriate F distribution, while the pairwise significance was tested by calculating the t statistic

$$t = \frac{\bar{X}_{td1} - \bar{X}_{td2}}{\hat{\sigma}_t \left[\frac{1}{n_2} + \frac{1}{n_2}\right]^{\frac{1}{2}}} \qquad (4.4.2)$$

and finding the significance from the t distribution.

4.5 The Experimental Results

Table 4.5.1 shows the results of the first scoring procedure as applied to the three tests. A summary of the distortions for each test is given in Table 4.5.2. The overwhelming point is that there are

	7 Digit Test					8 Digit Test				
	DISTORTION (t)					DISTORTION (t)				
	AV.		(1)	(2)	(3)	(4)	AV.		(1)	(2)	(3)	(4)
NOISE TEST	.29	(1)	X	1.86	2.00	2.71	.53	(1)	X	.37	.86	1.60
	.42	(2)	*	X	.14	.86	.56	(2)		X	.49	1.23
	.43	(3)	*		X	.71	.60	(3)			X	.74
	.48	(4)	**			X	.66	(4)				X
LPF TEST	.28	(1)	X	1.29	2.29	2.43	.55	(1)	X	1.36	1.98	2.22
	.37	(2)		X	1.00	1.14	.66	(2)		X	.62	.89
	.44	(3)	*		X	.14	.71	(3)	*		X	.24
	.45	(4)	**			X	.73	(4)	*			X
ADPCM TEST	.29	(1)	X	2.00	3.14	4.43	.56	(1)	X	1.36	2.22	3.09
	.43	(2)	*	X	1.14	2.43	.67	(2)		X	.86	1.73
	.51	(3)	**		X	1.29	.74	(3)	*		X	.86
	.60	(4)	**	**		X	.81	(4)	**			X

t LEVEL FOR SIGNIFICANCE FOR NO REJECTED DATA

* = Significance at .05

** = Significance at .01

TABLE 4.5.1 RESULTS OF UNSCREENED FIRST SCORING TESTS

TEST	DISTORTION			
	(1)	(2)	(3)	(4)
ADDITIVE NOISE	NONE	10db SNR	8db SNR	5db SNR
LOW PASS FILTER	NONE	2.4 kHz	1.8 kHz	1.2 kHz
ADPCM	NONE	24 KBPS	16 KBPS	8 KBPS

TABLE 4.5.2 DISTORTION LEVELS FOR THE TEST DIGITS ON THE THREE COMMUNICABILITY TESTS

very few significant results using this scoring scheme. The major problem here turns out to be the subject variations. Some subjects are so "bad" that they get practically no strings correct. Others are so "good" that they never miss. It was hence decided to screen out subjects whose average error rate was outside the range .3 < error rate < .7. This left 10 subjects on the first test, 9 on the second, and 10 on the third. The results for this scoring is shown in Table 4.5.3. Clearly, this screening improves the results, with a large number of results significant at the .01 level. This same effect was found to hold for the other two scoring procedures.

Tables 4.5.4 and 4.5.5 show the results from the second and third scoring procedures. In these tests the subjects were screened exactly as for the first scoring procedure. Several results are clear from these two tests. First, both scoring procedures represent a considerable improvement over the first procedure, with the third procedure having a slight edge in significance. Second, the noise tests seem to have less overall effect (less significance) than either the low pass filter test, or the ADPCM test. Third, the 7 digit test seems to be generally more acceptable than the 8 digit test (higher significance levels for the same number of subjects).

4.6 Conclusions

The purpose of this study was to ascertain the usefulness and cost of the digit recall test as a communicability test for speech digitization systems. The overall results must be stated to be that:

1. For the rather severe variations in distortions used in this test, it was easily possible to differentiate between systems.

	7 Digit Test DISTORTION					8 Digit Test DISTORTION				
			(1)	(2)	(3)	(4)			(1)	(2)	(3)	(4)
NOISE TEST	.29	(1)	X	2.68	3.35	4.70	.50	(1)	X	1.49	2.42	2.80
	.41	(2)	**	X	.67	2.01	.58	(2)		X	.93	1.30
	.44	(3)	**		X	1.34	.63	(3)	*		X	.37
	.50	(4)	**	*		X	.65	(4)	**			X
LPF TEST	.28	(1)	X	2.01	4.70	6.71	.51	(1)	X	2.24	4.10	4.47
	.37	(2)	*	X	2.68	4.70	.63	(2)	*	X	1.86	2.24
	.49	(3)	**	**	X	2.02	.73	(3)	**	*	X	.37
	.58	(4)	**	**	*	X	.75	(4)	**	*		X
ADPCM TEST	.28	(1)	X	3.58	5.59	8.05	.54	(1)	X	2.61	4.29	5.22
	.44	(2)	**	X	2.01	4.47	.68	(2)	*	X	1.66	2.61
	.53	(3)	**	*	X	2.46	.77	(3)	**	*	X	.93
	.64	(4)	**	**	*	X	.82	(4)	**	*		X

t LEVEL FOR SIGNIFICANCE FOR NO REJECTED DATA

* Significance at .05
** Significance at .01

TABLE 4.5.3 RESULTS OF SCREENED FIRST SCORING TESTS

	7 Digit						8 Digit					
			DISTORTION						DISTORTION			
			(1)	(2)	(3)	(4)			(1)	(2)	(3)	(4)
NOISE TEST	.62	(1)	X	2.47	5.22	8.51	.84	(1)	X	3.84	5.11	6.14
	.72	(2)	**	X	2.75	6.04	.99	(2)	**	X	1.28	2.30
	.81	(3)	**	**	X	3.30	1.04	(3)	**		X	1.02
	.93	(4)	**	**	**	X	1.08	(4)	**	**		X
LPC TEST	.60	(1)	X	4.67	6.32	10.44	.82	(1)	X	5.37	7.19	8.55
	.77	(2)	**	X	1.65	5.77	1.03	(2)	**	X	1.79	3.58
	.83	(3)	**	*	X	4.12	1.10	(3)	**	*	X	1.79
	.98	(4)	**	**	**	X	1.17	(4)	**	**	*	X
ADPCM TEST	.58	(1)	X	4.39	6.87	8.79	.83	(1)	X	4.60	6.65	8.18
	.74	(2)	**	X	2.41	4.39	1.01	(2)	**	X	2.05	3.58
	.83	(3)	**	**	X	1.92	1.09	(3)	**	**	X	1.53
	.90	(4)	**	**	**	X	1.15	(4)	**	**	*	X

^{t} LEVEL FOR SIGNIFICANCE FOR NO REJECTED DATA

* Significance at .05

** Significance at .01

TABLE 4.5.4 RESULTS OF THE SCREENED TESTS USING THE SECOND SCORING METHOD

7 Digit DISTORTION		(1)	(2)	(3)	(4)
.53	(1)	X	3.85	7.42	10.44
.67	(2)	**	X	3.57	6.59
.80	(3)	**	**	X	3.02
.91	(4)	**	**	**	X

8 Digit DISTORTION		(1)	(2)	(3)	(4)
.63	(1)	X	4.86	7.67	9.46
.82	(2)	**	X	2.81	4.60
.93	(3)	**	**	X	1.79
1.00	(4)	**	**	*	X

.51	(1)	X	3.57	7.69	9.89
.64	(2)	**	X	4.12	6.32
.79	(3)	**	**	X	2.20
.87	(4)	**	**	**	X

.61	(1)	X	5.63	8.44	10.74
.83	(2)	**	X	2.81	5.11
.94	(3)	**	**	X	2.30
1.03	(4)	**	**	**	X

.52	(1)	X	4.94	7.69	9.61
.70	(2)	**	X	2.75	4.67
.80	(3)	**	**	X	1.96
.87	(4)	**	**	*	X

.60	(1)	X	5.63	8.69	10.74
.82	(2)	**	X	3.07	5.11
.94	(3)	**	**	X	2.05
1.02	(4)	**	**	**	X

t LEVEL FOR SIGNIFICANCE FOR NO REJECTED DATA

* = Significance at .05

** = Significance at .01

TABLE 4.5.5 RESULTS OF SCREENED TESTS USING THE THIRD SCORING METHOD

2. The cost of this test is quite high when compared to other speech quality and speech intelligibility tests.
3. There is great subject variability, indicating that results might be improved substantially by using a trained, well documented crew of listeners.
4. For this particular group of subjects, 7 digits seemed about right. Clearly, however, for some 7 was too many, while for others, 8 was too few.
5. The test is a very unpleasant test in which to participate.
6. The ability of digit recall tests to differentiate between systems which are closely matched for performance is limited, and would require considerable cost.

In summary, it may be said that, even though this type of communicability test can be argued to be more appropriate than subjective preference testing, and even though it is possible, as shown in this study, to differentiate among distorting systems, still the excessive cost of communicability testing required to obtain the desired significance levels makes these tests unattractive.

APPENDIX A

SPEECH ACCEPTABILITY EVALUATION AT DYNASTAT: THE DIAGNOSTIC ACCEPTABILITY MEASURE (DAM)

APPENDIX A

SPEECH ACCEPTABILITY EVALUATION AT DYNASTAT: THE DIAGNOSTIC ACCEPTABILITY MEASURE (DAM)

BACKGROUND

It is a matter of common observation that user acceptance of voice communications equipment depends on factors other than speech intelligibility. Although a high degree of intelligibility is generally a necessary condition, it is not a sufficient condition of user acceptance. But until recently, no generally satisfactory methods of evaluating the overall acceptability or "quality" of processed or transmitted speech has been available. Among the previously available methods, some are applicable only for certain types of speech signal degradation. Others are of limited reliability. Virtually none permits reliable system evaluation in absolute terms for the diversity of processing techniques and transmissions encountered in modern digital voice communications.

Under contract with the Defense Communications Agency, Dynastat recently undertook to fill the need that existed in the area of acceptability evaluation. The results of this effort included the Paired Acceptability Rating Method (PARM) and the Quality Acceptance Rating Test (QUART), both of which provide improved reliability of measurement on an absolute scale of acceptability. Having met the interim needs of the Narrow Band Voice Consortium, they also served as valuable research tools to clarify a number of crucial methodological issues and to indicate possible means of further refining the technology of speech evaluation.

Drawing on insights gained in the course of its contractual activities with PARM and QUART, Dynastat continued under its own auspices to further advance the technology of communication

system evaluation from the standpoint of overall speech acceptability. These efforts culminated in the Diagnostic Acceptability Measure (DAM).

THE DIAGNOSTIC ACCEPTABILITY MEASURE

The Diagnostic Acceptability Measure combines direct (isometric) and indirect (parametric) approaches to acceptability evaluation by means of twenty-item system rating form.* Ten of the items on the form are concerned with the acceptability-related perceptual qualities of the speech signal, itself. Seven items are concerned with the perceptual qualities of the background. Three items are concerned with the perceived intelligibility, pleasantness, and overall acceptability of the total effect. The descriptors used to define the various perceptual qualities are the end products of an extensive program of research concerned with the nature of these qualities and with the development of a precise vocabulary for characterizing them.

The results of further research have indicated that listener's perceptions of modern digital voice communication systems and diverse forms of laboratory degradation can be exhaustively characterized in terms of six elementary perceived

---

* The isometric approach requires the listener to provide a direct subjective assessment of the acceptability of a sample speech transmission. The parametric approach requires the listener to evaluate the sample transmission with respect to various perceived characteristics or qualities (e.g., noisiness) independently of his individual effective reactions to these qualities. Hence, the parametric approach tends to minimize the sampling error associated with individual differences in "taste." The individual who does not personally place a high valuation on a particular speech quality may nevertheless provide information of use in predicting the typical individual's acceptance of speech characterized by a given degree of that perceptual quality.

qualities of the signal and three perceived qualities of the background. Measures of these elementary qualities are obtained by various combinations of rating scale data.

In accordance with the above research results, DAM rating data are presently analyzed to yield system diagnoses with respect to the nine perceptual qualities indicated in Table 1. The contribution of each of these qualities to the listener's acceptance reaction has been determined, so that each diagnostic score can be expressed in terms of the level of acceptability a system would be accorded if it were deficient with respect only to the single perceptual quality involved. Expressed in this way, the pattern of diagnostic scores reflects the relative contribution of each perceptual quality to the acceptability of the system, and permits the system developer to concentrate on the perceived characteristics of his system which are most detrimental to its acceptance.

The application of multiple, nonlinear regression techniques to a set of diagnostic scores permits the derivation of supplementary, parametric estimates of intelligibility, pleasantness, and acceptability, which can be combined with direct, or isometric rating data to yield highly reliable and valid estimates of all three of these properties. For practical purposes of system evaluation, however, parametric predictions are presently provided only for acceptability.

To permit comparisons with the results of tests previously conducted with PARM, DAM acceptability results are transformed to their PARM equivalents. A transformation of judged intelligibility results permits estimates of equivalent DRT total scores.

Rigorous procedures for monitoring and screening of listening crew members contribute significantly to the reliability of DAM results.

TABLE I. SYSTEM CHARACTERISTICS EVALUATED BY DAM

SIGNAL QUALITIES

Diagnostic Scale	Typical Descriptor	Exemplar	Intrinsic Effect On Acceptability
SF	Fluttering	Interrupted or Amplitude Modulated Speech	Moderate
SH	Thin	High Pass Speech	Mild
SD	Rasping	Peak Clipped Speech	Severe
SL	Muffled	Low Pass Speech	Mild
SI	Interrupted	Packetized Speech with "Glitches"	Moderate
SN	Nasal	2.4K bps Systems	Moderate

BACKGROUND QUALITIES

Diagnostic Scale	Typical Descriptor	Exemplar	Intrinsic Effect On Acceptability
BN	Hissing	Noise Masked Speech	Moderate
BB	Buzzing	Tandemmed Digital Systems	Moderate
BF	Babbling	Narrow Band Systems with Errors	Severe
BR*	Echoic	Multipath Transmission	?

TOTAL EFFECT

Scale
Intelligibility
Pleasantness
Acceptability

* Tentative scale, still under investigation.

Speaker differences are relatively small with DAM, particularly within sexes. Depending on the purposes of the investigator, however, the use of more then one speaker may be appropriate.

The speech materials used for purposes of DAM evaluations consist of 12 phonemically controlled sentences, spoken by each of the desired number of speakers. Approximately one minute total running time is required for each speaker.

Figure 1 shows the standard format in which DAM results are reported. Presented first are the basic diagnostic scores and their standard errors. Each diagnostic score represents one estimate of the acceptability rating the system being evaluated would receive if it were deficient only with respect to the corresponding perceptual quality. Summary scores, representing the combined effects of signal qualities and background qualities, respectively are also shown. Gross scores relating to acceptability, judged pleasantness and judged intelligibility are shown in the bottom half of the figure.

Isometric scores are based only on direct ratings of the respective characteristics.

Parametric scores are based on predictions of acceptability from combined diagnostic scores for signal quality and combined diagnostic scores for background quality.

Composite scores for acceptability are based on isometric scores for acceptability, parametric scores for acceptability, and on composite ratings of pleasantness and intelligibility.

DIAGNOSTIC ACCEPTABILITY MEASURE RESULTS

CUSTOMER: VOICE SYSTEMS INC. CONDITION: SPEECH PROCESSOR A-1 DATE RUN: 1 SEPT 1976

NUMBER OF LISTENERS: 10

SPEAKER(S): CH

SIGNAL QUALITY SCORES

FACTOR	DESCRIPTOR	EXEMPLAR	MEAN	SE
SF	FLUTTERING	INTERRUPTED SPEECH	78.4	3.2
SH	THIN	HIGH PASS SPEECH	83.3	2.3
SD	HARSH	PEAK CLIPPED SPEECH	74.4	1.6
SL	MUFFLED	LOW PASS SPEECH	80.0	3.0
SI	INTERRUPTED	PACKETIZED SPEECH WITH GLITCHES	87.4	3.4
SN	NASAL	2.4K bps SYSTEMS	80.0	1.2
TOTAL SIGNAL QUALITY			31.1	1.8

BACKGROUND QUALITY SCORES

FACTOR	DESCRIPTOR	EXEMPLAR	MEAN	SE
BN	HISSING	NOISE MASKED SPEECH	83.7	1.2
BF	BABBLING	NARROW BAND SYSTEMS WITH ERRORS	89.1	2.5
BB	BUZZING	TANDEMMED DIGITAL SYSTEMS	66.9	1.2
TOTAL BACKGROUND QUALITY			82.4	1.0

TOTAL EFFECT SCORES

RATING DIMENSION	ISOMETRIC SCORES MEAN	ISOMETRIC SCORES SE	PARAMETRIC SCORE MEAN	PARAMETRIC SCORE SE	COMPOSITE SCORE* MEAN	COMPOSITE SCORE* SE		EQUIVALENTS MEAN	EQUIVALENTS SE
INTELLIGIBILITY	55.3	2.4					(DRT)	84.1	2.0
PLEASANTNESS	31.0	2.3						----	---
ACCEPTABILITY	45.2	2.0	46.3	1.5	46.4	1.2	(PARM)	54.1	1.2

*Composite acceptability score based on combination of isometric scores for pleasantness, intelligibility and acceptability and parametric score for acceptability.

Dynastat, Inc.
2704 Rio Grande, Suite 4
Austin, TX 78705

Figure 1. Specimen Printout of DAM Results

Equivalent PARM scores and Equivalent DRT scores are currently obtained by simple linear regression techniques applied to composite acceptability scores and isometric intelligibility ratings, respectively. However, it is expected that more precise estimates of DRT scores will be obtained in the future through the application of multiple prediction techniques to the DAM diagnostic scores. Fig. 2 shows the correlation between DAM acceptability scores (composite) and PARM test results for a sample of modern digital voice communication systems. Fig. 3 shows the correlation between isometric intelligibility ratings and DRT total scores.

DAM evaluations have been performed on an extremely broad sample of state-of-the-art narrow band and broad band digital voice communication systems. Norms for various conditions of speech/noise ratio, band restriction, and other simple forms of signal degradation have also been established. These normative data provide Dynastat with truly unique capability for detailed, useful interpretation of DAM for future experimental systems or conditions. Research, contemplated and in progress, will serve to expand DAM's range of application and provide norms for yet to-be-encountered processing techniques and transmission conditions.

For further information regarding the technical aspects of the DAM and on the evaluation services Dynastat offers with it please contact:

Dr. William D. Voiers
Dynastat, Inc.
2704 Rio Grande, Suite 4
Austin, Texas 78705

Phone: (512) 476-4797

Administrative or contractual information relating to Dynastat's services with the DAM may be obtained from Mr. Ira L. Panzer at the same address and phone number.

APPENDIX B

DERIVATION OF THE PROBABILITY DENSITY FUNCTION FOR THE STUDENTIZED RANGE STATISTIC

APPENDIX B

DERIVATION OF THE PROBABILITY DENSITY FUNCTION FOR THE STUDENTIZED RANGE STATISTIC

From Figure 3.1, let

$$X_{max} = \alpha \qquad X_{min} = \beta \tag{B.1}$$

Then

$$Z = \alpha - \beta \tag{B.2}$$

and

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{\alpha\beta}(\xi, \xi - z)\, d\xi \tag{B.3}$$

as shown in Papoulis [B.1].

Now the correlative distribution of α and β is

$$\begin{aligned} F_{\alpha,\beta}(x,y) &= P(\alpha \le x,\ \beta \le y) \\ &= P(\{x_i \le x \text{ i}\} \cap \{x_j \le y \text{ for at least one j}\} \\ &= P(\{x_i \le x \text{ i}\} \cap \{x_i \ge y \text{ i}\}^C) \\ &= \begin{cases} F_X^k(x) - [F_X(x) - F_X(y)]^k & x > y \\ 0 & x \le y \end{cases} \end{aligned}$$

Then the joint probability density of α and β is

$$f_{\alpha,\beta}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{\alpha,\beta}(x,y) \qquad (B.5)$$

$$= \begin{cases} k(k-1)(f_X(x)-F_X(y))^{k-2} f_X(x) f_X(y) & x > y \\ 0 & x \leq y \end{cases}$$

Thus

$$f_Z(z) = \begin{cases} k(k-1) \int_{-\infty}^{+\infty} [F_X(w)-F_X(w-z)]^{k-2} f_X(w) f_X(w-z)\,dw & z > 0 \\ 0 & z < 0 \end{cases}$$

REFERENCES

B.1 Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1965.

APPENDIX C

MINICOMPUTER BASED
DIGITAL SIGNAL PROCESSING LABORATORY

A minicomputer-based Digital Signal Processing Laboratory has been under construction at Georgia Tech since August 1973. It is now an extensive hardware-software complex dedicated to research and instruction in many digital signal processing and minicomputer related areas. This appendix describes briefly the elements of this system.

The system is based upon three minicomputers, an Eclipse 5230 with 64K of 16-bit memory, and a NOVA 830 with 64K of 16-bit memory in the Research Lab, and a NOVA 820 with 32K of 16-bit memory in the Student Lab. The uses of these computers are numerous and diverse. Hence, the various hardware and software components of the system will be presented separately.

THE RESEARCH COMPUTERS

A block diagram of the basic research computer facility is shown in Figure 1. Included in this section are only those peripherals which are used by many applications. A full set of peripherals are listed in Table 1.

The computational power for the system is supplied by two groups of the Eclipse 5230, which has 64K of 16-bit semiconductor memory (+ CACHE), a floating-point processor, hardware multiply divided, a memory management unit, and writable control storage (for microprogramming the processor), and by one ground of the NOVA 830, which has a floating-point processor, hardware multiply-divide, a memory management unit, and 64K of 1 μsec 16-bit memory. Bulk storage is supplied by three discs. The main disc is a 192 M Byte moving head drive shared gy the Eclipse and the NOVA 830. Each of the

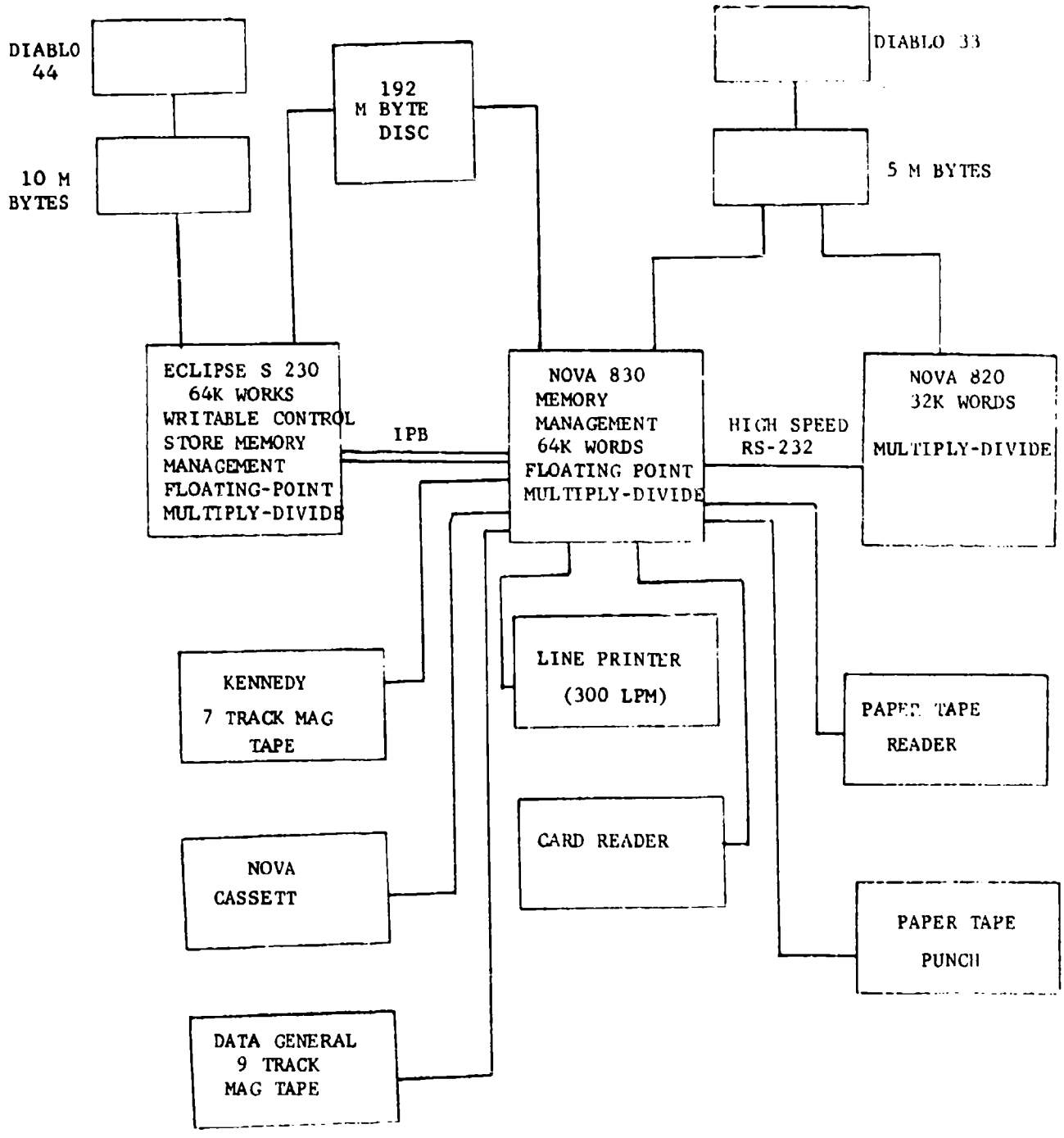


FIGURE 1

The Basic System for the Research Laboratory

TABLE 1

I/O DEVICES ON THE NOVA 830 I/O BUSS

DATA GENERAL INTERFACES

Diablo 33 disc controller (5 M bytes)
Diablo 44 disc controller (10 M bytes)
NOVA cassette controller
Real time clock
Floating-point arithmetic unit
Memory management
Data General mag tape controller
RS-232 interface at 9600 baud
RS-232 interface at 1200 baud
Inter-processor buss
Comtal video system interface

INTERFACES CONSTRUCTED AT GEORGIA TECH

Programmable sampling clock
RS-232 variable baud clock
Joy stick interface
Light pen interface
Button box interface
RS-232 interface (2)
16 bit double buffered D-to-A
10 bit single buffered D-to-A (4)
A-to-D/sample and hold/analog multiplexer
Ampex analog tape deck control
Revox analog tape deck control
Crown analog tape deck control
Kennedy 7-track digital tape interface
Line printer interface
Card reader interface
Paper tape reader interface
Programmable stack (256 words)
Quality test interface
Universal card tester interface
Time-of-day and date clock
Control card testing interface

other two disc units is of the moving head type, and each has one fixed and one removable pack. The Diablo model 44 disc has 10 M byte capacity, and is used by the Eclipse alone. The Diablo model 33 has 5 M byte capacity, and is shared by the NOVA 830 and the NOVA 820 (instructional) computers. Additional bulk storage is supplied by two tape units, a NOVA cassette tape and a 7-track digital unit (a 9-track unit is on order from Data General). The cassette is standard Data General peripheral, while the 7-track was interfaced at Georgia Tech.

Additional general purpose devices include a card reader, a line printer, a paper tape reader, and a paper tape punch. These units were all interfaced at Georgia Tech.

The foreground of the NOVA 830 is used a a general peripheral control ground for sharing the scarce peripherals. Most all of the general purpose and special purpose peripherals in the system are interfaced to the NOVA 830 (see Table 1), and this ground accesses all the other grounds on the other computers in the system to access these peripherals.

THE GRAPHICS SUBSYSTEM

One of the major design criteria for this system was a high level of high speed graphical interaction between the user and the computer. Figure 2 shows the hardware associated with the graphical subsystem.

This system supports many types of graphical interaction. First, it supports line printer graphics both in the axis-graph mode and in the X-Y-Z mode for picture reproduction. Second, the Tektronix

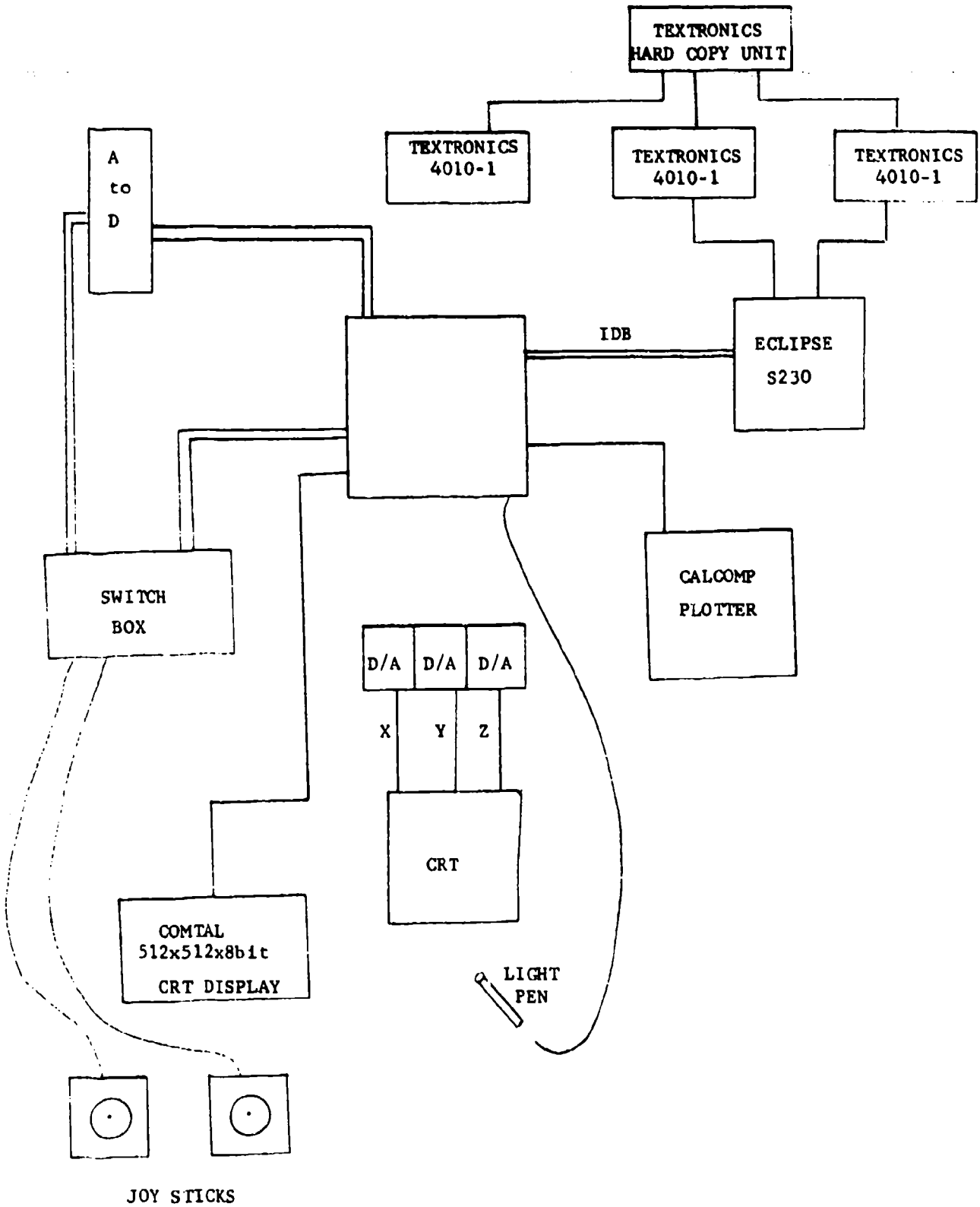


FIGURE 2

The NOVA 830 Graphical Subsystem

4010 graphical unit gives storage type vector graphics at 9600 baud and cross hair feedback interaction. Third, refresh graphics is supplied by driving X-Y-Z CRT's directly from 3 of the D-to-A's. A light pen (built at Georgia Tech), along with two joy sticks, 3 button boxes, and two potentiometers give interactior in the refresh mode. Fourth, a CALCOMP incremental plotter (interfaced at Georgia Tech), gives hard copy capability in the vector and character modes. Last, a Comtal video processor gives X-Y-Z CRT support on a 512x512 display with eight bits resolution.

THE AUDIO SUBSYSTEM

A diagram of the audio subsystem is given in Figure 3. This subsystem was constructed as an aid to interactive speech processing.

The whole system is centered on a patch bay located with the NOVA 830. This patch bay gives the user great flexibility in interconnecting the individual audio components.

Data acquisition is handled through a 12-bit (10 μsec) A-to-D with an 8-channel analog multiplexer on its input. Data playback is handled by a 16-bit double buffered D-to-A. The sampling rate on these two units is controlled by a programmable clock. Four additional channels of 8 bit D-to-A's form single buffered analog outputs. The entire data acquisition and playback system was built at Georgia Tech.

Four analog tape drives are available for use with the system. Two of these, a Crown 800 and a Revox tape drive, are interfaced so they may be controlled by the computer. The Crown interface allows the positioning of the tape to any desired position (within tape stretch). Either of the two Ampex drives may be used under computer

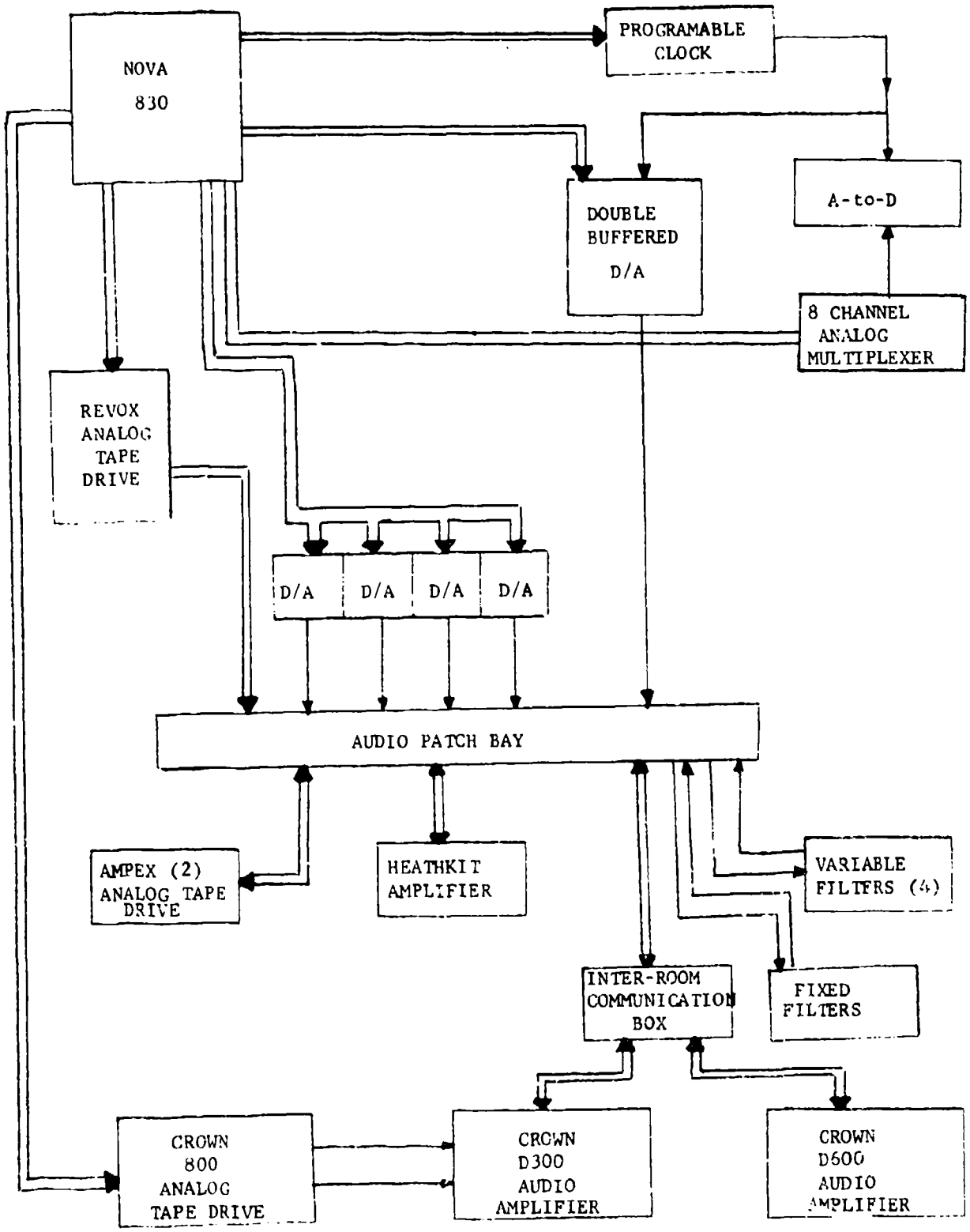


FIGURE 3

The Audio Subsystem on the NOVA 830

control in place of the Revox.

Four variable filters and three audio amplifiers are also available for use with this system.

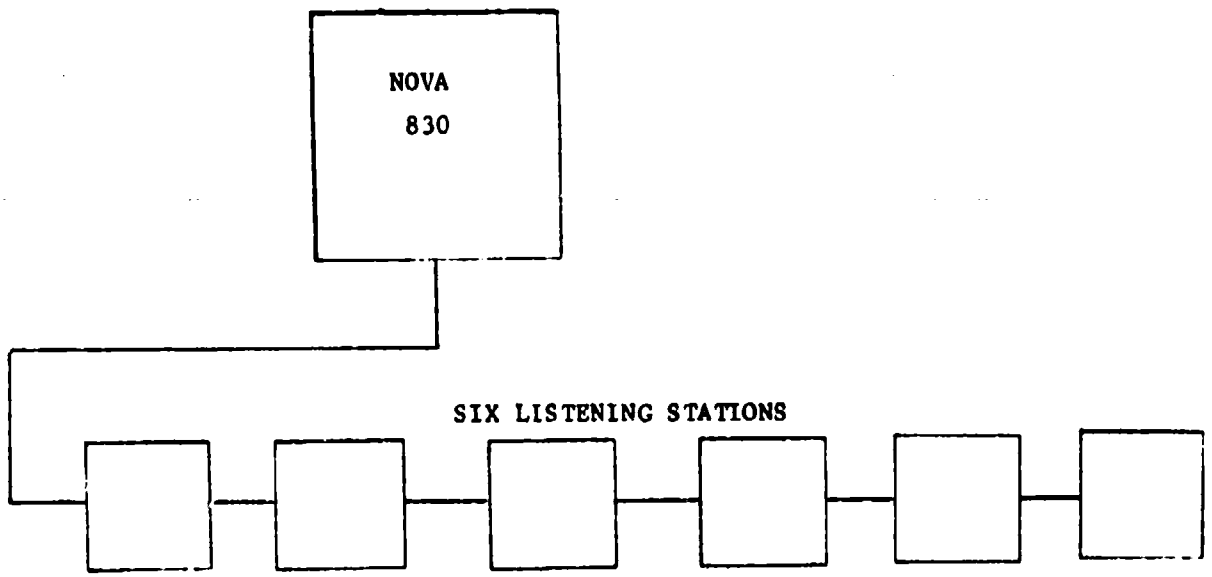
SPEECH QUALITY TEST SUBSYSTEM

The speech quality test subsystem depicted in Figure 4 is designed for the automated control of subjective quality tests. The subsystem consists of six stations, located in a separate speech quality laboratory and controlled by the NOVA 830 computer. Each of the stations has a CRT, 15 response buttons, a "read" button, ear phones, and a volume control for each ear. The computer interface can read the buttons at any station, clear and set the ready flip flop, and, using a software character generator, display messages to the subjects on the CRT's.

This quality system has several distinct advantages over a non-automated system. First, it eliminates much of the hand work on data reduction. Second, it allows on-line statistical analysis. Last, it allows the subjective test to reconfigure itself based on the subject responses.

THE OPTICAL DATA PROCESSING SUBSYSTEM

A diagram of the optical data processing facility is given in Figure 5. This subsystem has three components. The first component, the "picture acquisition" component, consists of a Micro NOVA Microcomputer (in Dr. William Rhodes' laboratory) which controls an electro-mechanical scanner. This equipment is still under development. Second, the Micro NOVA also controls an optical data digitizer for picture acquisition. The third component in this system is the



EACH STATION: CRT
15 BUTTONS
READY BUTTON
EAR PHONES & VOLUME CONTROL

FIGURE 4

The Speech Quality Testing Subsystem

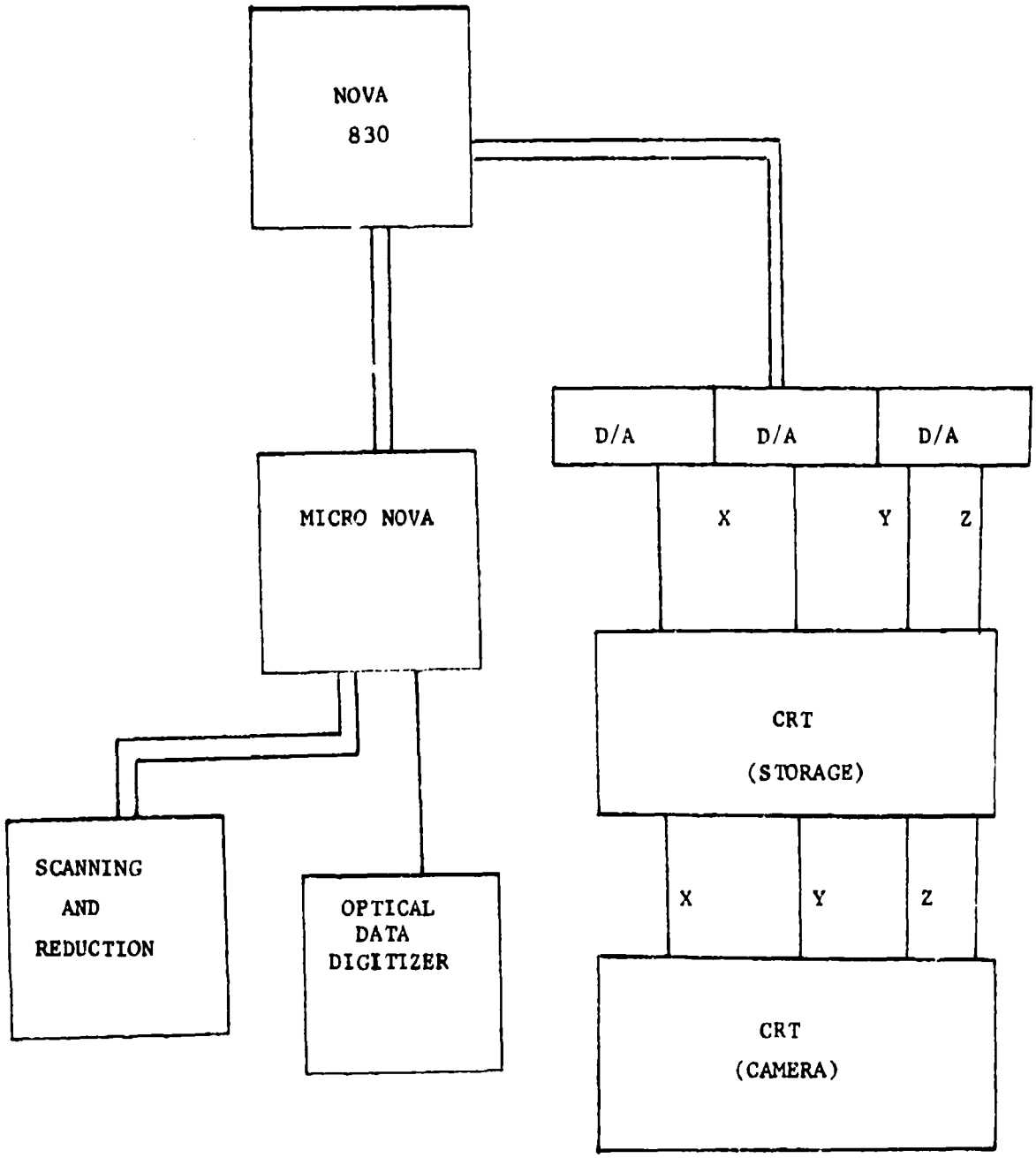


FIGURE 5

The Optical Data Processing Subsystem

"picture playback" facility. This facility consists of 3 D-to-A's and two CRT scopes. One CRT is of the storage type, and allows quick viewing of the pictures being displayed. The second CRT is equipped with a scope camera. The interchangeable backs on this camera allow the production of either polaroid or 120 roll film pictures. The Comtal video system can also be used to produce pictures.

THE COMPUTER NETWORK SUBSYSTEM

A "star" computer network is currently under development in the digital signal processing laboratory. The basic hardware for this system is shown in Figure 6. The NOVA 830 communicates with the Eclipse through an interprocessor buss (IPD), and with several other computers through high speed, variable baud rate, RS-232 standard, asynchronous, serial interfaces. These RS-232 interfaces were designed and built at Georgia Tech, and are capable of speeds up to 152K baud.

The hardware for this system exists and is tested. The software is currently under development.

THE UNIVERSAL CARD TESTER AND THE HARDWARE PHILOSOPHY

One of the most important subsystems of the digital signal processing laboratory is the universal card tester. To understand how this is used, it is important to understand the hardware philosophy of the laboratory. Most of the hardware constructed in the laboratory is constructed in prebuilt chassis. Each chassis contains 40 56-pin connectors. The computer I/O buss enters each chassis and is split into 3 sub-busses, called the "data buss," the "control buss," and the "address buss." If this is not the final chassis on the daisy chain, the busses are regrouped, and extended to the next chassis.

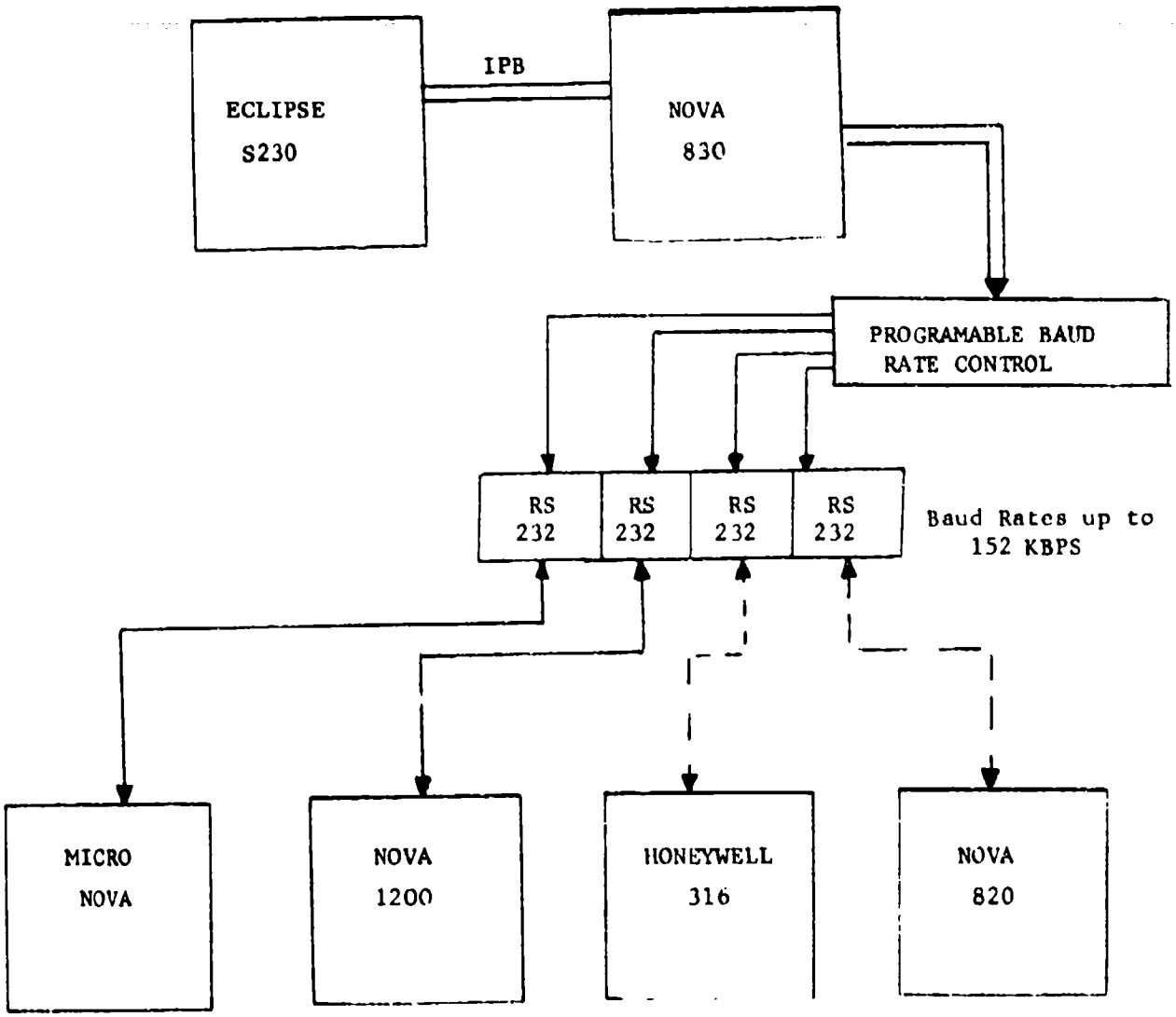


FIGURE 6

The Computer Network Subsystem

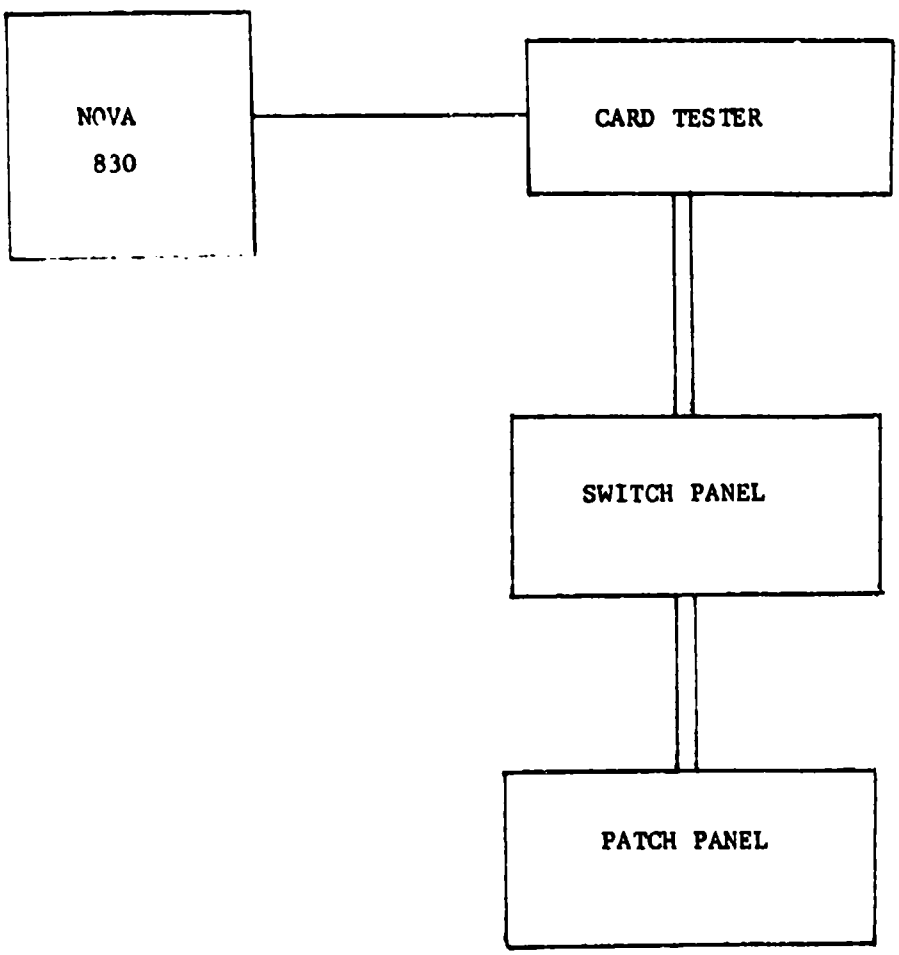


FIGURE 7

The Universal Card Tester

The hardware interfaces constructed in the chassis are mostly constructed from pre-designed printed circuit boards. A list of the PC cards available for interface construction is given in Table 2. Most interfaces consist of using some set of "standard" cards with, perhaps, some additional construction.

The main problem in hardware construction, therefore, is in building and testing the "standard" cards, often with semi-skilled labor. This is the purpose of the universal card tester.

A diagram of the universal card tester is given in Figure 7. The tester has a switch panel, a patch panel, and a single "standard" 56-pin connector as an "input," and "output," or as an "external." Each pin has a parallel connection to the patch panel for external connection. The computer can read or write individual bits to any pin position. Hence, any desired input/output sequence can be presented to a card being tested, and the results can be read back by the computer.

The software package associated with the card tester allows the user to test and debug any of the standard cards. In addition, a special card allows the testing of individual integrated circuit chips.

THE BASIC INSTRUCTIONAL COMPUTER (NOVA 820)

The NOVA 820 computer and its associated peripherals forms a computer and signal processing facility dedicated to student activities. These activities mainly include several laboratories associated with course and student project work. The hardware is configured so as to allow maximum utilization of the software developed in the research laboratory.

TABLE 2

STANDARD PC CARDS USED IN THE MODULAR CONSTRUCTION SYSTEM

CARD NAME	PURPOSE
Single Address	Address decode
Dual Address	Address decode
Control	Interrupt control
Input buffer	16 bit input buffer
Output buffer	16 bit output buffer
DMA	Direct memory access control
Counter	16 bit up/down counter
Memory	256x256 bit high speed memory (43 msec)
RS-232 (1)	High speed serial converter
RS-232 (2)	Medium speed serial converter
M6800 CPU	Micro-processor CPU
M6800 Memory (1)	Micro-processor memory (4K Ram)
M6800 Memory (2)	Micro-processor memory (4K RAM, 4K ERROR)
M6800 Buffer	Micro-processor buffer
M6800 Control	Micro-processor interrupt control
Kluge	General purpose

Figure 8 shows the basic NOVA 820 computer system and Table 3 gives a list of peripherals. The CPU has 32K of 800 nsec memory and a hardware multiply-divide unit. Bulk storage is formed by two moving-head disc drives totaling 5 M bytes of storage. These discs are shared with the NOVA 830, and communication between the processors is maintained on a high speed RS-232 port.

Many of the peripherals have been constructed so as to be identical, from a computer command viewpoint, to those on the research facility. Hence, the D-to-A's, the double buffered D-to-A's, the A-to-D, the A-to-D 8-channel analog multiplexer, and the programmable clock all utilize the same commands as their counterparts on the NOVA 830. These peripherals give the NOVA 820 a similar audio and refresh graphics capability to the NOVA 830.

Interactive graphics on the NOVA 820 is handled by a M6800 control plasma terminal designed to look like a Tektronixs 4010. Hence, all the graphics packages developed for the NOVA 830 will run on the NOVA 820.

THE MICRO-COMPUTER SUBSYSTEM (M6800)

One of the most important developments in modern control technology has been the development of the micro-processor. The micro-processor subsystem of the student (NOVA 820) laboratory was developed with three purposes:

1. To develop a micro-processor board set for use as a general interfacing tool.
2. To develop a hardware interface between NOVA 820

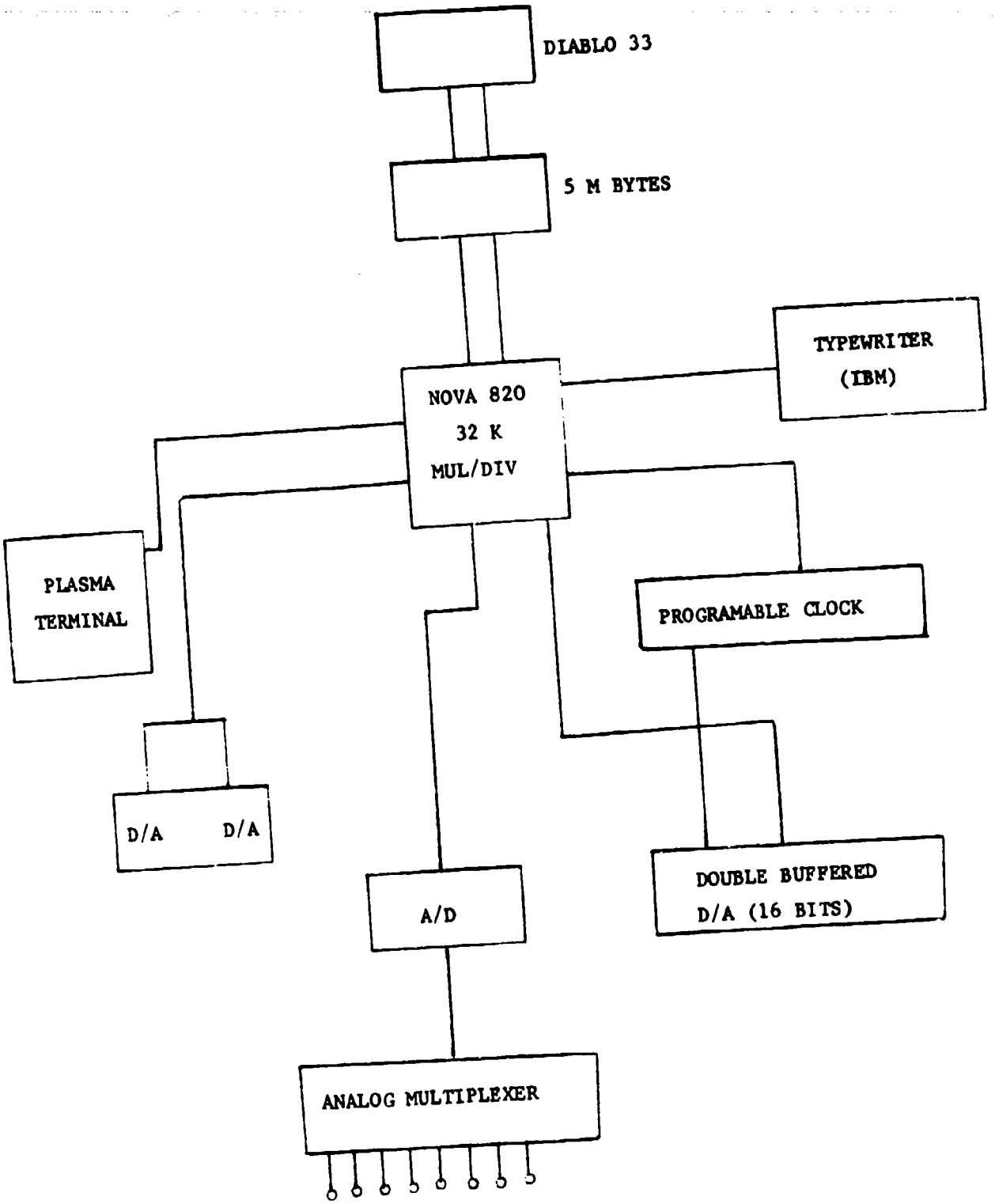


FIGURE 8

The NOVA 820 Basic System

TABLE 3

I/O DEVICES ON THE NOVA 820 I/O BUSS

DATA GENERAL INTERFACES

Diablo 33 disc controller
RS-232 interface at 1200 baud
Inter-processor buss

INTERFACES CONSTRUCTED AT GEORGIA TECH

Programmable sampling clock
Light pen interface
16 bit double buffered D-to-A
10 bit single buffered D-to-A (4)
A-to-D/sample and hold/analog multiplexer
Line printer/M6800 input interface
M6800 Micro-computer CPU
M6800 4K memory module (2)
M6800 control and communication interface
Plasma display interface

and a micro-processor and to develop software for the NOVA 820 which allow simple, interactive software development for the microprocessor.

2. To develop software for the micro-processor to do the graphics and character generation tasks related to the plasma scope.

All three of these purposes have been accomplished. Future goals for the subsystem include the addition of another 8 bit micro-processor board *8080A) and the development of a system based on the new Data General 16 bit micro-processor.

A diagram of the hardware associated with the micro-processor is shown in Figure 9. Through a general interface to the micro-processor's buss, the NOVA 820 can completely control the micro-processor and load and examine the micro-processor memory. Through a standard interrupt interface, the NOVA 820 can communicate with the micro-processor as it would any other peripheral. This environment allows great flexibility in the use of the micro-processor.

The micro-processor itself has 8K of 8 bit, 1 msec memory, an interrupt I/0 port, and a 16 bit I/0 buffer. Expansion of the hardware and software for this subsystem is continuing.

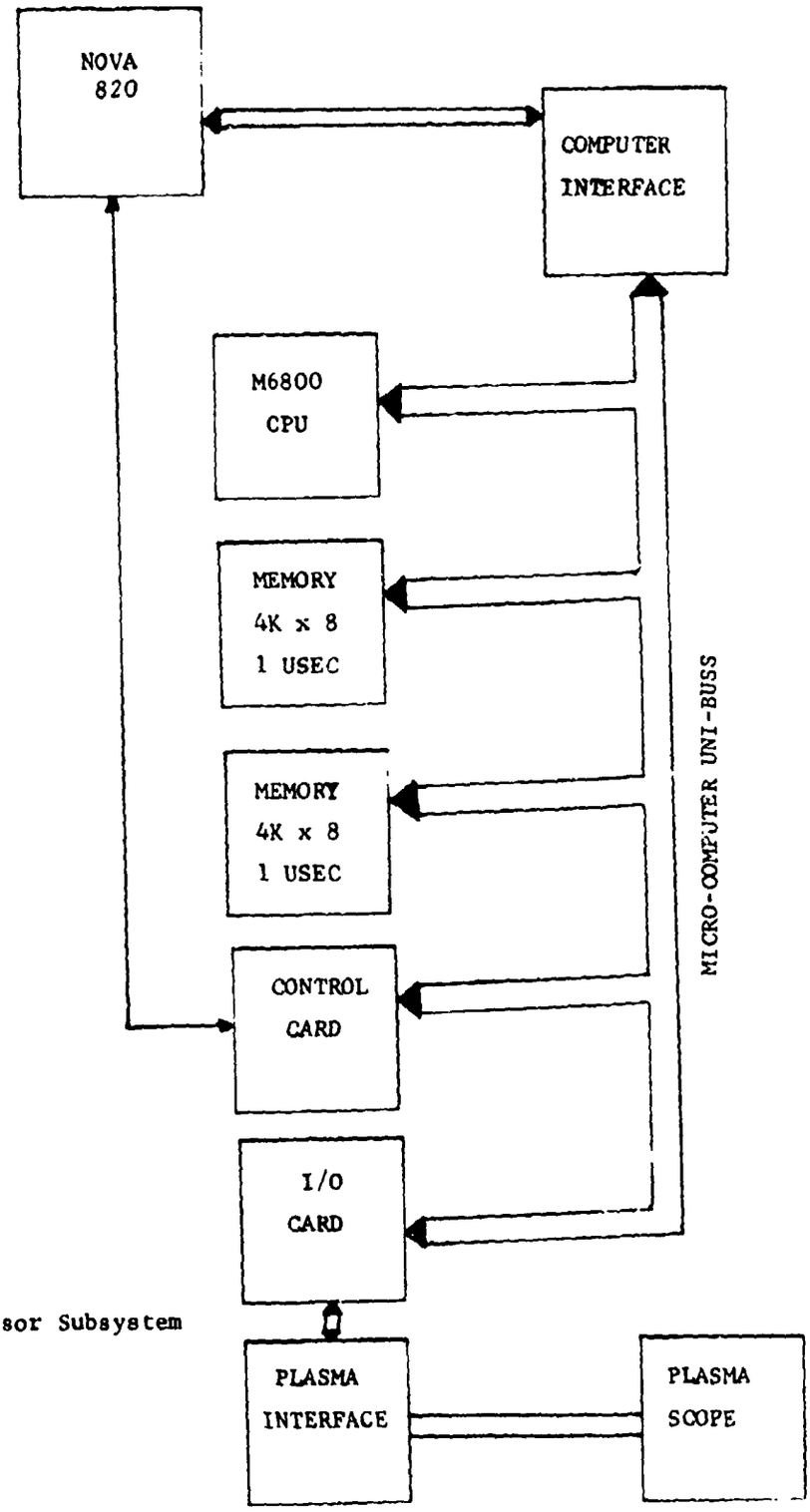


FIGURE 9

The Micro-Processor Subsystem

APPENDIX D

SOFTWARE SUMMARY

PROGRAM NAME: ACONT
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	G	INPUT STARTING ADDRESS FROM TTY
R	G	DATA IS REAL(ASSUME INTEGER OTHERWISE)
O	L	OUTPUT (CONTIGUOUS) FILE -- MUST COME FIRST

PURPOSE
TO CONCATENATE A SET OF CONTIGIOUS FILES INTO A SINGLE OUTP

PROGRAM NAME: ACONTS
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
R	G	DATA IS REAL--ASSUMED INTEGER OTHERWISE
O	L	CONTIGIOUS OUTPUT FILE

PURPOSE
TO CONCATENATE A SET OF CONTIGIOUS INPUT FILES OF INTEGRAL N
OF CYLINDERS INTO A SINGLE OUTPUT FILE

PROGRAM NAME: ADPCM
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
P	L	PITCH FILE
I	L	INPUT FILE (SPEECH)
O	L	OUTPUT FILE (SPEECH)
C	L	FEEDBACK COEFFIENT FILE
X	L	QUANTIZED ERROR OUTPUT FILE
E	L	ERROR OUTPUT FILE
M	L	MULTIPLIER OUTPUT FILE
D	L	DATA FILE
L	L	LISTING FILE

PURPOSE
TO SIMULATE GENERAL ADPCM SYSTEMS. SYSTEM IS CONFIGURED BY D
AND INPUT/OUTPUT FILES(EG. IF A /P FILE IS PRESENT, A PITCH S
ERROR CORRECTION IS DONE)

PROGRAM NAME: CPITCH
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
O	L	OUTPUT PITCH FILE

PURPOSE
TO CREATE A CONSTANT PITCH CONTOUR.

PROGRAM NAME: DECK
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
P	C	PLAY
R	C	RECORD
F	C	FAST FORWARD
B	C	FAST BACKWARD
C	C	USE CROWN INSTEAD OF AMPEX

PURPOSE
ANALOGUE TAPE DRIVE CONTROL PROGRAM.

PROGRAM NAME: DCAADIN
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: GENERAL

PURPOSE:
THIS IS AN INTERACTIVE PROGRAM FOR TRANSFERING DATA FROM IBM
SPEECH DATA TAPES, ORGINATING AT DCA, TO DATA GENERAL CONFIG

FILES
THE PROGRAM IS INTERACTIVE AND SELF EXPLANITORY

PROGRAM NAME: DCAAV
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: GENERAL

PURPOSE
THIS PROGRAM COMPUTES THE AVERAGE OF MANY OBJECTIVE MEASURES COMPUTED BY OBJETIVE AND OBJ2. ITS PURPOSE IS TO GET AN OVERALL MEASURE FROM MANY SINGLE WINDOWED ERRORS.

PROGRAM NAME: DCATAPEIN
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: GENERAL

PURPOSE
THIS IS AN INTERACTIVE PROGRAM TO TRANSFER AN IBM 9 TRACK TAPE CODED IN EBCDIC TO AN ASCII FILE ON RDOS FILE STRUCTURE

PROGRAM NAME: DATAMAKE
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	L	INPUT INSTRUCTION FILE
O	L	OUTPUT INSTRUCTION FILE
D	L	DATA FILE

PURPOSE
TO MAKE A NEW DATA FILE FOR THE SYSTEMTIC TESTING OF ANY SYSTEM.

PROGRAM NAME: DATASTART
LANGUAGE: FORT
CATEGORY: GENERAL

PURPOSE
INTERACTIVE PROGRAM FOR CREATING CONTROL FILE FOR DATAMAKE.

PROGRAM NAME: DFDP
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	R A	INPUT DATA FILE (OPTIONAL)
O	R A	OUTPUT FILTER COEFFICIENTS
M	R A	MAGNITUDE SPECTRUM (OPTIONAL
P	R A	PHASE SPECTRUM (OPTIONAL)

PURPOSE
DESIGNS DIGITAL FILTERS

PROGRAM NAME: DFD
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	R A	INPUT DATA FILE (OPTIONAL)
O	R A	OUTPUT FILTER COEFFICIENTS
M	R A	MAGNITUDE SPECTRUM (OPTIONAL
P	R A	PHASE SPECTRUM (OPTIONAL)

PURPOSE
DESIGNS DIGITAL FILTERS

PROGRAM NAME: DOWN
LANGUAGE: FORT
CATEGORY: SPEECH

PURPOSE
TO DROP LOWER ORDER BITS, AND/OR DROP EVERY OTHER OR 2 OUT O
OR ------ BITS TO REDUCE SAMPLING FREQUENCY

PROGRAM NAME: FILTER
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
R	L	RESULT FILE
D	L	DATA FILE

PURPOSE
GENERAL CANOICAL FORM DIGITAL FILTER PROGRAM

PROGRAM NAME. FNORM
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
R	L	RESULT FILE
D	L	DATA FILE

PURPOSE
TO NOMALIZE A FLOATING POINT FILE.

PROGRAM NAME: FFILTER
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
R	L	RESULT
D	L	DATA FILE

PURPOSE
FORGROUND VERSION OF FILTER.

PROGRAM NAME: FILMPY
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
O	R A	OUTPUT FILTER COEFF
M	R A	MAGNITUDE SPECTRUM (OPTIONAL)
P	R A	PHASE SPECTRUM (OPTIONAL)

PURPOSE
PUTS TOGETHER ANY NUMBER OF DIGITAL FILTERS TO MAKE
ONE FILTER (CASCADE). INPUT FILTER FILES HAVE NO SWITHCES.

PROGRAM NAME. FILPLT
LANGUAGE FORT
CATEGORY GENERAL

PURPOSE
F-SWAP PROGRAM FOR DFDP

PROGRAM NAME: GOGO
LANGUAGE FORT
CATEGORY: SPEECH

PURPOSE
TO INITIALIZE THE CLOCK AND A/D CHANNEL

PROGRAM NAME: HEAR
LANGUAGE ASM
CATEGORY SPEECH

SWITCH	TYPE	PURPOSE
*	L	SEE *

PURPOSE
TO SAMPLE INPUT ANALOGUE DATA

* SWITCH DETERMINES SIZE OF SAMPLE INCYLINDERS
A=1, B=2, ETC.

PROGRAM NAME: HLPD
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT SPEECH DATA
P	L	OUTPUT PITCH DATA
D	L	DATA FILE
L	L	LISTING FILE

PURPOSE
HARD LIMITED AUTOCORRELATION PITCH DETECTOR.

PROGRAM NAME: HIRE
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	I A	INTEGER SPEECH INPUT FILE
O	I A	INTEGER IMPULSE RESPONSE OUTPUT
P	R A	DATA FILE (OPTIONAL)
L	L	LISTING (OPTIONAL)

PURPOSE
HOMOMORPHIC IMPULSE RESPONSE EXTRACTOR.

PROGRAM NAME: LPC

LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT SPEECH FILE
C	L	COEF. FILE
K	L	PARCOR COEF. FILE
R	L	AUTO. FILE
D	L	DATA FILE
L	L	LISTING FILE

PURPOSE
BASIC BLOCK SYNCHRONOUS AUTOCORRELATION/TOEPLITZ VOCODER TRA

PROGRAM NAME: LPR
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
A	LOCAL	AREA FUNCTIONS
K	LOCAL	PARCOR COEFFICIENTS
C	LOCAL	FFEDBACK COEFFICIENTS
O	LOCAL	FEEDBAK COEFFICIENTS
D	LOCAL	BATCH (DATA) CONTROL FILE
R	LOCAL	AUTOCORRELATION COEFFICIENTS
P	LOCAL	PITCH FILE
L	LOCAL	LISTING FILE
X	LOCAL	EXCITATION OUTPUT FILE

PURPOSE
THIS IS A GENERAL PURPOSE LPC RECEIVER PROGRAM. IT RECONFIGU ITSELF DEPENDING ON WHAT FILES APPEAR IN ITS INPUT COMMAND LINE. IF ITS "X" LINES ARE COMPILIED, THE PROGRAM CAN ADD SEVERAL DISTORTIONS TO THE OUTPUT SPEECH, INCLUDING UNIFORM BANDWIDTH DISTORTION AND UNIFORM FREQUENCY DISTORTION. IT MA THUS BE USED TO CORRECT HELIUM SPEECH OR INSTALL CONTROLLED DISTORTIONS ON THE OUTPUT.

PROGRAM NAME: LOOK
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	L	DATA FILE

PURPOSE
INTERACTIVE GRAPHICS INTERPRETER WHICH ALLOWS UP TO EIGHT PI.
BASED ON UP TO 4 FILES ON THE 4010 GRAPHICS TERMINAL.

PROGRAM NAME: MBPD
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
A	I A	UNFILTERED SPEECH INPUT
B	I A	50-100HZ FILTERED SPEECH
C	I A	100-200HZ FILTERED SPEECH
2D	I A	200-400HZ FILTERED SPEECH
E	I A	400-800HZ FILTERED SPEECH
I	I A	DATA FILE INPUT (OPTIONAL)
P	R A	PITCH CONTOUR OUTPUT
L	R	AVERAGE LEVEL INPUT (FROM MBPWR)

PURPOSE
MULTI BAND PITCH DETECTOR

PROGRAM NAME: MBPLOT
LANGUAGE: FORT
CATEGORY: SPEECH

PURPOSE
"F-SWAP"PROGRAM FOR USE WITH MBPD

PROGRAM NAME: MBPWR
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
A	I A	UNFILTERED SPEECH INPUT

```
B          I A     50-100HZ FILTERED SPEECH INPUT
C          I A     100-200HZ FILTERED SPEECH INPUT
D          I A     200-400HZ FILTERED SPEECH INPUT
E          I A     400-800HZ FILTERED SPEECH INPUT
O          I A     LEVEL OUTPUT FILE
```

PURPOSE
AVERAGE MAGNITUDE LEVEL FOR MBPD

PROGRAM NAME: NORM
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
R	L	RESULT FILE
D	L	DATA FILE

PURPOSE
TO NORMALIZE AN INTEGER FILE

```
C
C
C     PROGRAM NAME:        OBJECTIVE
C     LANGUAGE:            FORTRAN
C     DATE:                 6/ 9/77
C     AUTHOR:              T. P. BARNWELL
C     CATEGORY:            SPEECH
C
C
C     SWITCH    TYPE    PURPOSE
C
C     M         LOCAL   MASTER FILE
C     S         LOCAL   SLAVE FILE
C     D         LOCAL   BATCH (DATA) FILE
C     L         LOCAL.  LISTING FILE
C
C
C     PURPOSE
C       TO COMPUTE THE GAIN WEIGHTED AND NON GAIN WEIGHTED SPECTRAL
C       DISTANCE METRIC BETWEEN TWO SPECTRUM FILES. THE SPECTRUM
C       FILES ARE NORMALLY GENERATED BY LPC , PCEP, HIRE, OR SPCANA.
C
C
C       ------------------------------------
```

PROGRAM NAME: OBJ2
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTH R: T. P. BARNWELL
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
M	LOCAL	MASTER FILE
S	LOCAL	SLAVE FILE
D	LOCAL	BATCH (DATA) FILE
L	LOCAL	LISTING FILE

PURPOSE
TO COMPUTE THE GAIN WEIGHTED AND NON GAIN WEIGHTED NON--SPECTRAL DISTANCE METRIC BETWEEN TWO SPECTRUM FILES. THE NON-SPECTRUM FILES ARE NORMALLY GENERATED BY LPC , PCEP, HIRE, OR SPCANA.

PROGRAM NAME: PCEP
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
D	LOCAL	BATCH (DATA) CONTROL FILE
A	LOCAL	OUTPUT CEPSTRUM FROM A
M	LOCAL	MASTER INPUT
S	LOCAL	SLAVE INPUT(B)
B	LOCAL	OUTPUT CEPSTRUM FROM B
L	LOCAL	LISTING FILE
W	LOCAL	INPUT (ASCII) WINDOW (FIR FILTER) FUNCTION
Z	LOCAL	BINARY POINT BY POINT METRIC

PURPOSE
THIS IS A GENERAL PURPOS CEPSTRAL COMPARE PROGRAM. IT ALLOWS USER TO COMPARE ANY REGION OF THE OPOSING CEPSTRUMS AFTER AN WINDOW FUNCTION HAS BEEN APPLIED. THIS ALLOWS THE PROGRAM TO USED FOR BOTH SPECTRAL ENVELOP AND EXCITATION COMPARISONS

PROGRAM NAME: PDISTORT
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: SPEECH

PURPOSE:
THIS PROGRAM IS USED TO SYSTEMATICALLY DISTORT PITCH CONTOUR
THE DISTORTION IS A CONSTANT RISE OR FALL IN THE PITCH PERIO
THE DISTORTION ONLY OCCURES IN VOICED SEGMENTS, AND THE PROG
IS INTERACTIVE.

PROGRAM NAME: PTGTC
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
F	L	PITCH FILE
I	L	INPUT SPEECH
F	L	INPUT FILTERED SPEECH

PURPOSE
TO HAND PAINT A PITCH CONTOUR FOR TESTING.

PROGRAM NAME: PCHECH
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
M	L	INPUT STATISTICS FILE
T	L	OUTPUT STATISTICS FILE
D	L	DATA FILE
A	L	ADD ON HISTIGRAM IN
O	L	ADD ON HISTOGRAM OUT
L	L	LISTING

PURPOSE
TO CHECK THE OUTPUT OF A PITCH PERIOD ESTIMATOR AGAINST
A HAN PAINTER PITCH CONTOUR.

PROGRAM NAME: PRNT
LANGUAGE: FORT
CATEGORY: GENERAL

PURPOSE
TO PRINT A PROGRAM WITH FILE NAME AND DATE

PROGRAM NAME: SCALE
LANGUAGE: FORT
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
R	L	RESULT FILE

PURPOSE
TO SCALE A DATA FILE FOR FILTER

PROGRAM NAME: SF
LANGUAGE: FORT
CATEGORY: GENERAL

SWITCH	TYPE	PURPOSE
I	L	INPUT FILE
D	L	DATA FILE
R	L	RESULT FILE
C	L	COEF. FILE

PURPOSE
TIME VARYING DIGITAL FILTER PROGRAM

PROGRAM NAME: SPCANA
LANGUAGE: FORTRAN
DATE: 6/ 9/77
AUTHOR: T. P. BARNWELL
CATEGORY: SPEECH

SWITCH	TYPE	PURPOSE
I	LOCAL	INPUT FILE

```
O          LOCAL   OUTPUT SPECTRUM
D          LOCAL   BATCH (DATA) CONTROL FILE
L          LOCAL   LOG SPECTRUM OUTPUT FILE


PURPOSE
  THIS IS A GENERAL PURPOSE SPECTRUM ANALYSIS PROGRAM DESIGNED
  TO DO CEPSTRUM OR LPC DECONVOLVED SPECTRUM.

     ------------------------------------


PROGRAM NAME:       ZCPD
LANGUAGE:           FORT
CATEGORY:           SPEECH


SWITCH    TYPE    PURPOSE

I         L       INPUT FILE
P         L       OUTPUT PITCH CONTOUR
D         L       DATA FILE (OPTIONAL


PURPOSE
ZERO CROSSING PITCH DETECTOR

     ------------------------------------
```